LOS ALAMOS SCIENCE AND TECHNOLOGY MAGAZINE

NOVEMBER 2013

1663

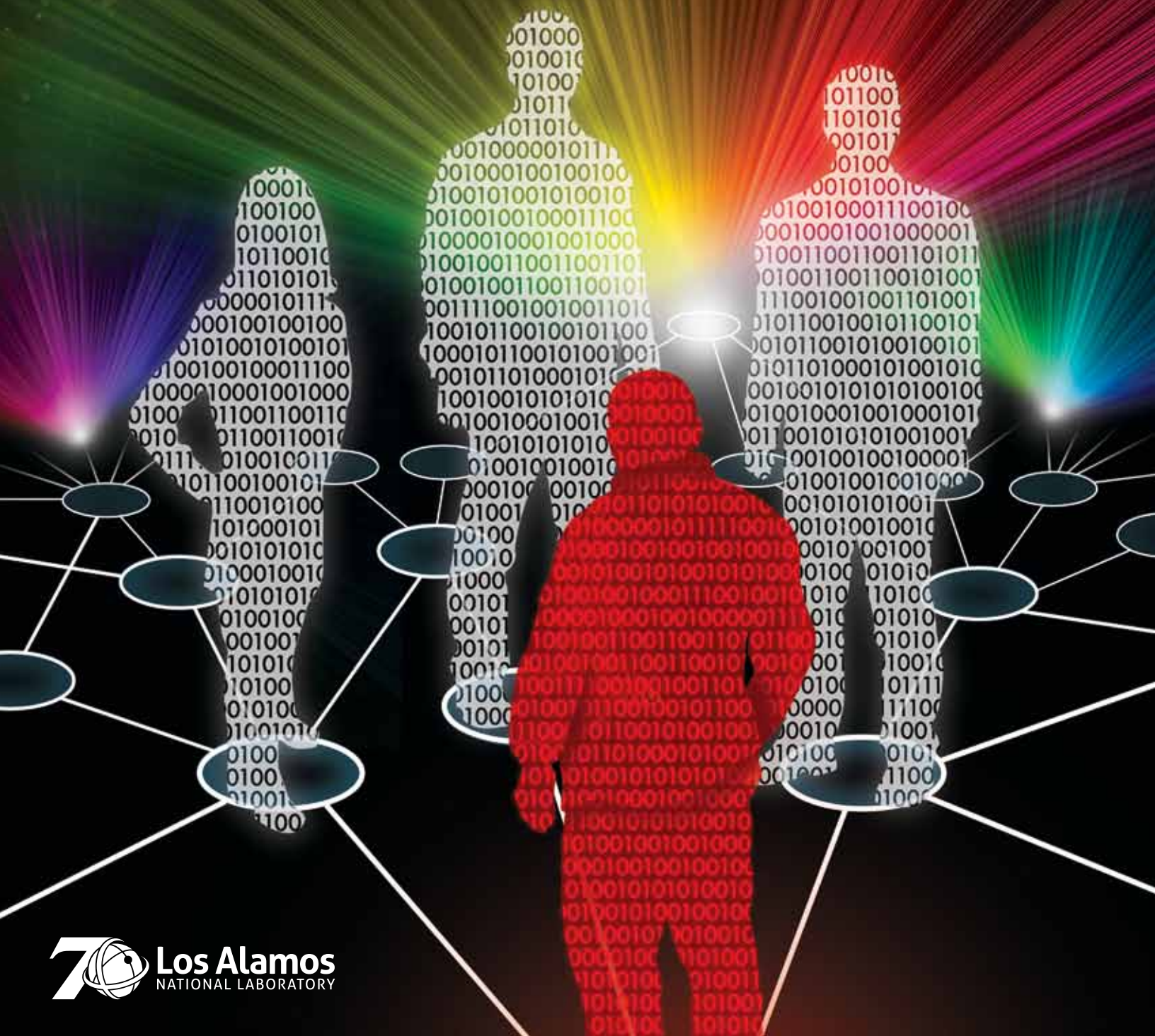
Twitter Vision

Human Genome 10 Years Later

Super-strange Superconductors

Good News from Glaciers

CYBER INTRUDER ALERT



Los Alamos NATIONAL LABORATORY

1663

LOS ALAMOS SCIENCE AND TECHNOLOGY MAGAZINE

About Our Name:
During World War II, all that the outside world knew of Los Alamos and its top-secret laboratory was the mailing address—P. O. Box 1663, Santa Fe, New Mexico. That box number, still part of our address, symbolizes our historic role in the nation's service.

About the LDRD Logo:
Laboratory Directed Research and Development (LDRD) is a competitive, internal program by which Los Alamos National Laboratory is authorized by Congress to invest in research and development that is both highly innovative and vital to our national interests. Whenever *1663* reports on research that received support from LDRD, this logo appears at the end of the article.

About the Cover:
With cyber attacks on American enterprises becoming increasingly common, cyber security experts at Los Alamos are deploying their latest and greatest defense, a software program called PathScan. When hackers invade a network, they move from one computer to another, creating a path through the network as they search for valuable data and ever-greater levels of access. PathScan, as its name implies, scans for such illicit paths. Unlike existing firewalls, which provide protection against unauthorized entry, PathScan catches intruders on the inside—but before they can do any damage.

Los Alamos Firsts

The First Cell Sorter

Imagine, as Los Alamos physicist Mack Fulwyler did in the mid-1960s, an instrument that could find the proverbial needle in a haystack, only the needles were certain living cells, and the haystack was a solution containing millions of other cells. Fulwyler turned imagination into reality and patented his "cell separator" in 1965. Cell biology hasn't been the same since.

The story begins at Los Alamos Scientific Laboratory in the early 1950s, in the days of atmospheric nuclear weapons testing. A team of Laboratory physicists had joined with chemists and biologists to develop methods to test for the presence of radioactive fallout in people and food products. But in 1963, with the limited test ban treaty in force, they turned to other biomedical measurement challenges, including DNA analysis, cancer cell detection, and blood cell characterization.

The team had acquired a Coulter Counter, a recently developed instrument that could measure the size of single cells in solution for automated blood analysis. The instrument forced a sample of cells in solution to flow past an electric sensing region, one cell at a time. As each cell passed, an electronic signal changed by an amount that was proportional to the size of the cell. The distribution of cell sizes gave important information on the distribution of cell types in the sample, which could then be linked to a diagnosis.

In studying this instrument, Fulwyler had the idea to isolate specific cells from the large population of cells to confirm their identities. He could use the electronic signals from the size analysis to trigger a mechanism that would divert specific cells from the fast-moving sample stream (thousands of cells per second). He didn't know what that mechanism would be, but knew that mechanical methods to physically separate individual cells were too slow.

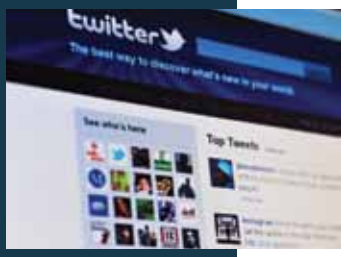
Then Fulwyler read a paper describing the invention of an ink-jet printing device in which a stream of ink was vibrated to break it down into droplets. The droplets were charged according to an electronic signal and subsequently redirected by an electric field. Fulwyler had his answer: he would marry the automated cell-size analysis concept with the fast ink-jet printing technology. Cells suspended in a conducting fluid would be partitioned into fine droplets after size analysis. Then a charge would be applied to the droplets containing the cells of interest, allowing those droplets to be pulled out of the stream for collection and further study. The first paper, describing the cell sorter and its positive results when sorting mixtures of human and mouse blood cells of different sizes, appeared in the top-tier journal *Science* in 1965 and a patent was issued.

In the 50 years or so since Fulwyler built the first cell separator, the descendents of that original instrument have become ubiquitous in research and clinical laboratories all over the world. The development of flow cytometry (cell measurement) with cell sorting transformed cellular biology from a descriptive, qualitative science to a quantitative science. With the addition of fluorescent dyes that chemically light up specific cell molecules, researchers have gained the ability to ask questions about cell function—what a cell does, rather than just what it looks like. Today, flow cytometry provides insight into the complex cellular and molecular mechanisms that underpin diseases such as cancer and AIDS and helps elucidate the role of individual cell types in health, disease, and treatment.

—Babetta Marrone

Director of the National Flow Cytometry Resource (2008–2013) at Los Alamos National Laboratory

IN THIS ISSUE



FEATURES

CREDIT: MATTHEW HOFFMAN/LANL

SPOTLIGHT

VISIONARY

how twitter is helping computers to see

HUMAN BEINGS ARE "VISIONARIES": we rely on our vision, almost to the exclusion of our other senses, to inform us and guide our interactions with the world. In broad strokes, our vision gives us the ability to recognize *what*—objects and people in our environment, things that are similar, things that are different—as well as the ability to determine *where*—where things are, where they were, and where they will be.

Computers are not visionaries. Even when configured with superb optical "eyes" and abundant "brain" power, most computers see poorly, in the sense that, in non-controlled environments, their ability to determine *what* or *where* is limited and inconsistent. A computer would have a very difficult time identifying everything that would catch a human's eye on a busy street. Furthermore, a computer optimized to recognize faces would likely fail miserably if asked to recognize a gun, or steer a car, or perform any visual task other than what it was optimized for.

Yet if current trends in computer technology, algorithm development, and data availability continue, computers will likely have excellent vision within the next five years. They will be able to process a complex visual scene quickly, accurately, and thoroughly, and will match or even outperform humans in most vision-specific tasks.

And helping to bring about that "see change" will be, of all things, Twitter, the social networking and microblogging service giant.

#Hey, look at this

It's a little odd that humanity's desire to blog is related to the pursuit of computer vision, but the connection exists because people often blog about what they see, and the computer has to be taught how to see. To recognize a cat, for example, the computer must first be given an image of a cat and told, "This is a cat." In fact, the computer needs to have seen thousands of cats—in all positions, in different environments, at various angles, and under arbitrary light conditions such as a visual blog site might provide—so that it can recognize a cat regardless of circumstances.

Humans also have to learn to see, only the process is innate and happens largely without supervision. Learning begins essentially at birth and continues unabated for several years. By the end of its first year, an infant will have observed about a petabyte (10^{15} bytes) worth of data, or enough to fill 100,000 ten-gigabyte thumb drives. No computer has ever come close to being trained on so large a data set. Indeed, scientists' inability to find a sufficiently large training set of annotated images has been a major stumbling block to realizing a sight-worthy computer.

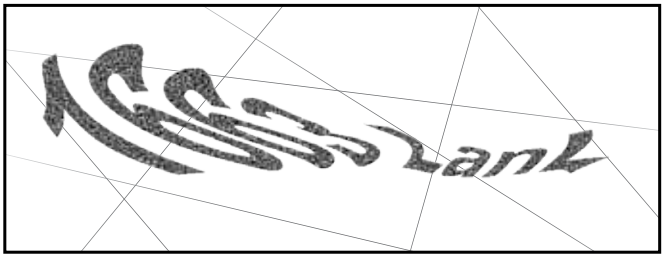
That changed on January 24, 2013, the day Twitter released Vine, an application that allows users to attach and send a short, six-second video tweet to followers. Vine has experienced exponential growth since then, and the amount of data sent collectively by Vine users has already topped 60 terabytes (60×10^{12} bytes). One couldn't ask for a better training set.

"Millions of people are sending six-second videos to each other, each annotated with a statement like, 'This is my cat playing,' or 'NYC's best taco,'" says Steven Brumby of the Los Alamos National Laboratory, the lead scientist in an effort to develop a computer's vision. "We can store the videos, then search the collection for "cat" and have ready access to hundreds of thousands of videos of cats. It's a remarkable resource."

Couple the training set with a supercomputer able to execute several trillion operations per second (teraflops), and the goal of computer vision comes into view.

"Our focus is to develop the basic mathematics and computer science underpinning computer vision," says Brumby. "I anticipate we'll have visually adept computers within two years, in part because Google, Amazon, Silicon Valley startups, and several big academic groups are all working to make computer vision happen. This is the holy grail."

It's the holy grail because a sighted computer would enable a range of vital applications, foremost being autonomous robots that can be used for defense, manufacturing, resource extraction, emergency disaster response, environmental assessment, etc. If able to evaluate its environment faithfully, a seeing computer would usher in an era of computer-controlled transportation—non-stop trucking, coordinated traffic flow, and autonomous minivans that pick the kids up from soccer. Furthermore, a seeing computer would be an exceptional personal assistant, one with full access to the Internet and its body of knowledge that could help you keep tabs on your loved ones and watch over the sick and elderly.



Computers have such a hard time interpreting a visual field that a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) has become a standard online security measure. Users are asked to extract information from a simple image and thus prove they are human beings.

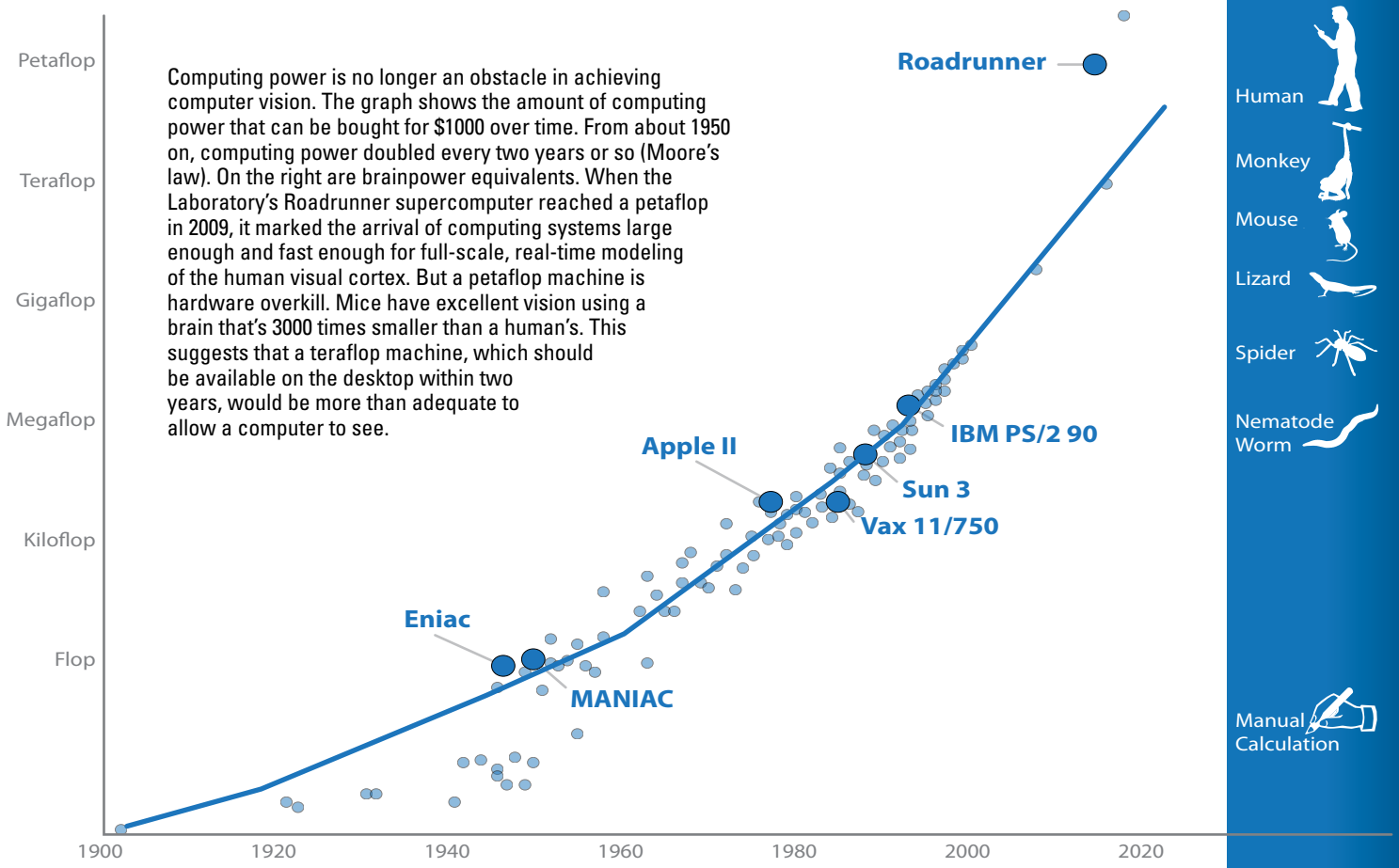
Unquestionably, a computer looking over your shoulder could be a good thing. But there are many who fear that a seeing computer will be the starting point of a sci-fi nightmare. Consider that the computer could use the camera in your smart phone or laptop, plus the network of traffic and security cameras that monitor essentially every street and alleyway of our cities, to identify you and the people around you and determine where you are and what you are doing. Private companies—for purely commercial reasons—are already beginning to master the technology for identifying and tracking consumers and their interests. Apart from raising issues of personal privacy, tracking can shift into surveillance, and the computer could be used to help achieve and sustain a police state.

But computers have been used and abused almost from the earliest days of computer applications. Today's powerful computers—the "predeprocessors" of which helped break Nazi communication codes during World War II—already monitor telecommunications and email in an effort to hunt down global terrorist groups. Despite this level of privacy intrusion, humanity has managed to thrive.

Let there be sight

How is a computer able to see? There is no simple answer, as currently the method pursued depends to a large extent on the visual task the computer will be performing, be it object recognition, event detection, video tracking, scene reconstruction, or something else. One area Los Alamos is pursuing is object recognition, basing its algorithms on models of how the human brain sees.

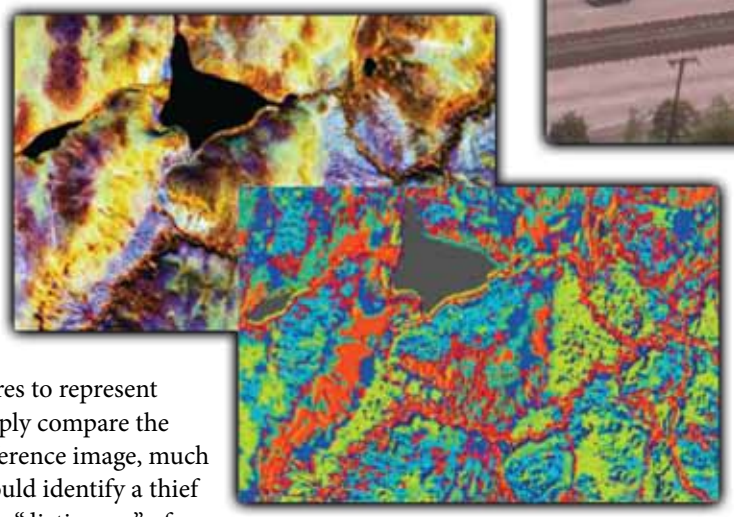
Briefly, object recognition starts by giving the computer a digital image that contains, say, a cat, and asking it to find all cats. The computer divvies the image, or portions of the image, into thousands, if not hundreds of thousands of tiny patches, with each patch being a tiny image perhaps 8 pixels by 8 pixels in size. The computer will try to represent, or duplicate, the information content of each patch by searching through a collection of patch-sized images that it has



Computing power is no longer an obstacle in achieving computer vision. The graph shows the amount of computing power that can be bought for $1000 over time. From about 1950 on, computing power doubled every two years or so (Moore's law). On the right are brainpower equivalents. When the Laboratory's Roadrunner supercomputer reached a petaflop in 2009, it marked the arrival of computing systems large enough and fast enough for full-scale, real-time modeling of the human visual cortex. But a petaflop machine is hardware overkill. Mice have excellent vision using a brain that's 3000 times smaller than a human's. This suggests that a teraflop machine, which should be available on the desktop within two years, would be more than adequate to allow a computer to see.

stored in memory, and selecting the ones that are a good match. The stored images are called features, and the collection of features is referred to as a dictionary.

The features at this first level of processing are simple—a line pitched at a certain angle, a blotch of color, etc. But once the computer has done its best to represent each patch by combining one or more features, it moves on to the second processing level. The small patches are grouped together into larger patches that cover a greater fraction of the object. These larger patches are then represented by combinations of features contained in a second-level dictionary. These features are more indicative of the cat than first-level features and might show, for example, the straight lines of its whiskers or the color and texture of its fur. After executing several similar levels of processing, the patches are large enough to include the entire object, and the computer has found a set of features that accurately represent the object(s) in the input image. The set is given to a classification program, which plays a multi-dimensional game of Twenty Questions before it says, "It's a cat."

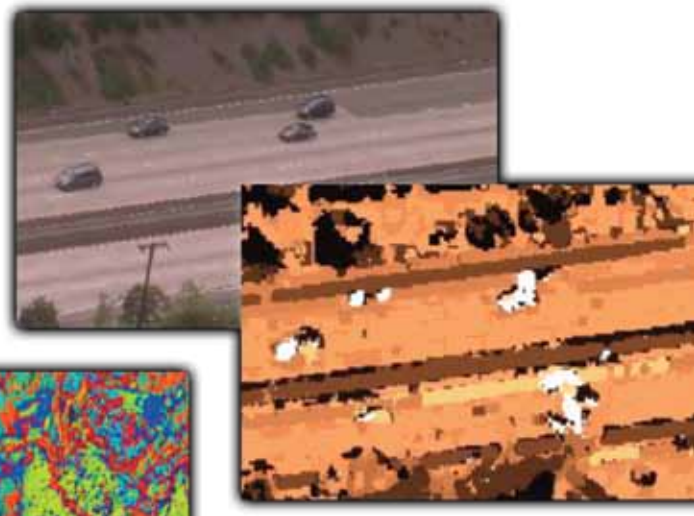


Why use features to represent objects, and not simply compare the entire image to a reference image, much the way a person would identify a thief by looking through a "dictionary" of mug shots? One reason is that the computer matches images by doing a pixel by pixel comparison and calculating a "distance parameter," with dissimilar images being farther apart.

Suppose there are two similar images of today's featured object, the cat, but in one, the cat's head is upright, while in the other, its tilted. A human would instantly recognize that its the same cat in both images, but the computer's pixel by pixel comparison would result in a large distance parameter, because a portion of the images don't line up. To obtain a closer match, the image dictionary would have to contain images of cats with their heads up, with their heads tilted, as well as every conceivable variation, so there would always have to be a stored image that aligns with the input image.

"Every object that we wanted the computer to recognize would need a similar portfolio of poses," says Brendt Wohlberg, a scientist working with Brumby. "The dictionary becomes extremely large and computationally very expensive to manipulate."

By breaking the image up into features, one can represent essentially any image by combinations of simpler images, much the way the entire English language can be constructed from combinations of just 26 letters.



Applications of a vision-capable computer include (left) detecting changing vegetation in multi-sensor satellite imagery and (right) detecting vehicles in aerial high-definition video. In each pairing, the upper image is the original observation, and the lower image is a computer-vision reconstruction.

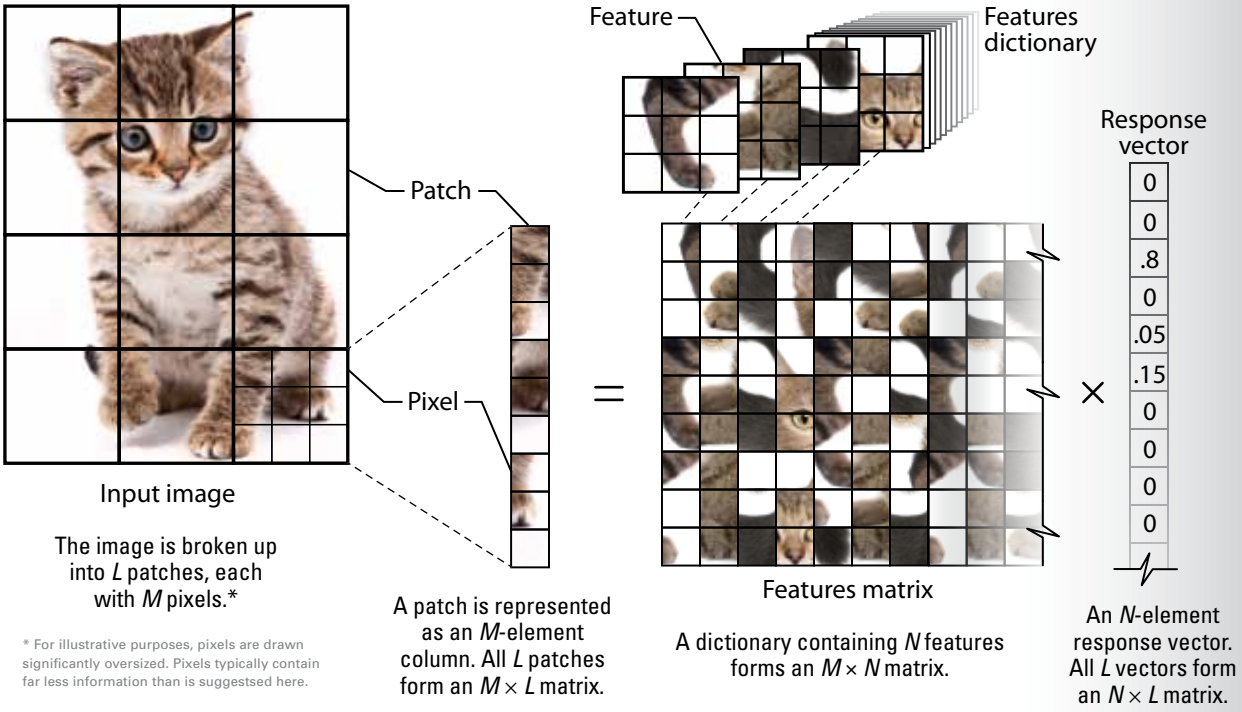
Processing an image, therefore, entails finding a way to represent hundreds of thousands of patches by combinations of just a few hundred features. Seeing in real time demands that the computer manipulate huge amounts of data—millions of pixels—very quickly. The processing rate needs to be teraflops or better. But the key to seeing lies in the features dictionaries, which need to have enough entries to represent the content of any patch. It's vital for the computer to have processed countless real-world images so that it can build its dictionary.

Sparse Training

"Natural images, or portions of a natural image, are generally not that complex," says Rick Chartrand, who shares lead investigator duties with Steven Brumby. "They can almost always be represented by a sparse representation of features, assuming one has a sufficient number of features to choose from."

Sparse representations, which figure heavily in the computer-vision algorithms being developed at Los Alamos, can be understood with a little help from a mixed drink. Consider a local bar, stocked with dozens of liquors, juices, sodas, waters and flavorful mixes—the full inventory forming a "liquids dictionary." Any drink can be made by mixing the elements in the dictionary together in various amounts. For example, a gin and tonic is made with 1 part gin, 3 parts tonic water, and 0 parts of every other liquid in the bar. Each drink is a sparse representation of the liquids dictionary, in that each is made by mixing together only a few of the dictionary's many elements.

Matrix equation for a single patch



* For illustrative purposes, pixels are drawn significantly oversized. Pixels typically contain far less information than is suggestsed here.

The computer recognizes an object by finding a sparse representation of it. The image is broken up into perhaps a hundred thousand tiny images called patches. Features are patch-sized color images that the computer can recall from a "dictionary" stored in memory. Like mixing a drink, just about any patch can be reproduced by mixing a few features together in various amounts. The amounts are encoded in an entity called a response vector.

The goal is to find the mix of features that accurately reproduces a patch, which entails finding the optimal reponse vector. Each patch or feature is made up of pixels, and by rearranging the pixels, one can represent the patches, the features dictionary, and the response vectors as matrices in a matrix equation. The optimal solution to the equation will be a set of sparse response vectors, each one a recipe for reproducing a patch as a sparse representation of the features dictionary.

In training a computer to see, a large set of, say, 500 images is processed simultaneously. At 100,000 patches per image, all patches form a matrix of 50 million columns, and the computer uses the same features dictionary to find the 50-million-column matrix of response vectors. Any of those solutions can be added to the dictionary to improve it. The computer can thus bootstrap and optimize both its dictionary and response vectors. The more images a computer processes, the better its features dictionaries and the better its sight.

On the corner of Twitter and Vine

The six-second Vine videos that can be attached to Twitter tweets are a computer-vision resource the likes of which the world has never encountered. The phone app is easy to use, and people have responded, filming and posting to the world microvideos of their cat at play, the friendly waiter at the restaurant, local street performers, or baby's first steps. When terrorist bombs exploded during the 2013 Boston Marathon, local Vine users posted video of the chaos almost instantaneously, while users everywhere took, then forwarded, video of television newscasts, spreading word of the disaster at an unprecedented pace. Access to the videos is free—they are public domain and any Twitter user can download them—making what happened in Boston available for public scrutiny and analysis.

When Los Alamos cosmologist and computer sensei Mike Warren, who also works with Brumby and Wohlberg, heard about the Vine release, he immediately recognized the potential to create a unique resource for vision research. Utilizing a prototype storage system (initially developed to archive astronomy data and the results of supercomputer simulations), he wrote software to download and archive the videos. The stream of data he started collecting in the early spring has since become a deluge. Warren estimates that during peak Vine usage this summer he was collecting more than a million videos per day.

A quick perusal of the data reveals an unrivaled training set. For example, searching the tweets for videos annotated with the word "cat" finds more than 250,000 videos, most with at least one cat in it. (A similar search finds more than 400,000 videos of dogs, apparently the Vine user's BFF). Selecting videos based on their dominant color, like green or blue, reveals thousands of short films showing grass or sky. Stills from the videos can be used for training, or the video themselves can be used to teach the computer to detect motion.

"Watching 24 hours a day, it would take 12 years to view the video we have now. That's an amount of information that rivals everything an 18-year-old has ever seen or heard," said Warren.

Where it stands

Life at a national nuclear security laboratory is a little different, in that security is a priority and, one way or another, affects every process and procedure. The institutional supercomputers that will be taught to see were ill-equipped to receive an ocean of unknown, unverified data downloaded from the Web, and system engineers and cybersecurity experts have yet to resolve the myriad of throughput and security issues. Only a tiny fraction of Vine data has been processed, and a remarkably patient Warren waits for whatever changes that need to be made to be made.

Undeterred, Brumby's team is continuing to explore different algorithms that will process more data faster and achieve a higher level of recognition fidelity. They are also refining methods that enable a computer to search through an enormous data set unsupervised, so that it learns on its own which features to extract to best help it identify objects.

Steven Brumby with a Vine backdrop filtered by the keyword "blue."

Will a computer ever truly be able to see? If sight is simply extracting information from light, then computers are already seeing, and Brumby and his team can be viewed as high-powered ophthalmologists working to make computers see better. But seeing is often tied to awareness, to interpreting the light-based information so as to understand the world around us. While cognizant computers are a staple of science fiction, they are at present not part of the real world. It's doubtful a computer will ever see things the way we do. But whether human-like or not, computers will attain excellent vision within our lifetime, and the world around us will be forever changed. LDRD

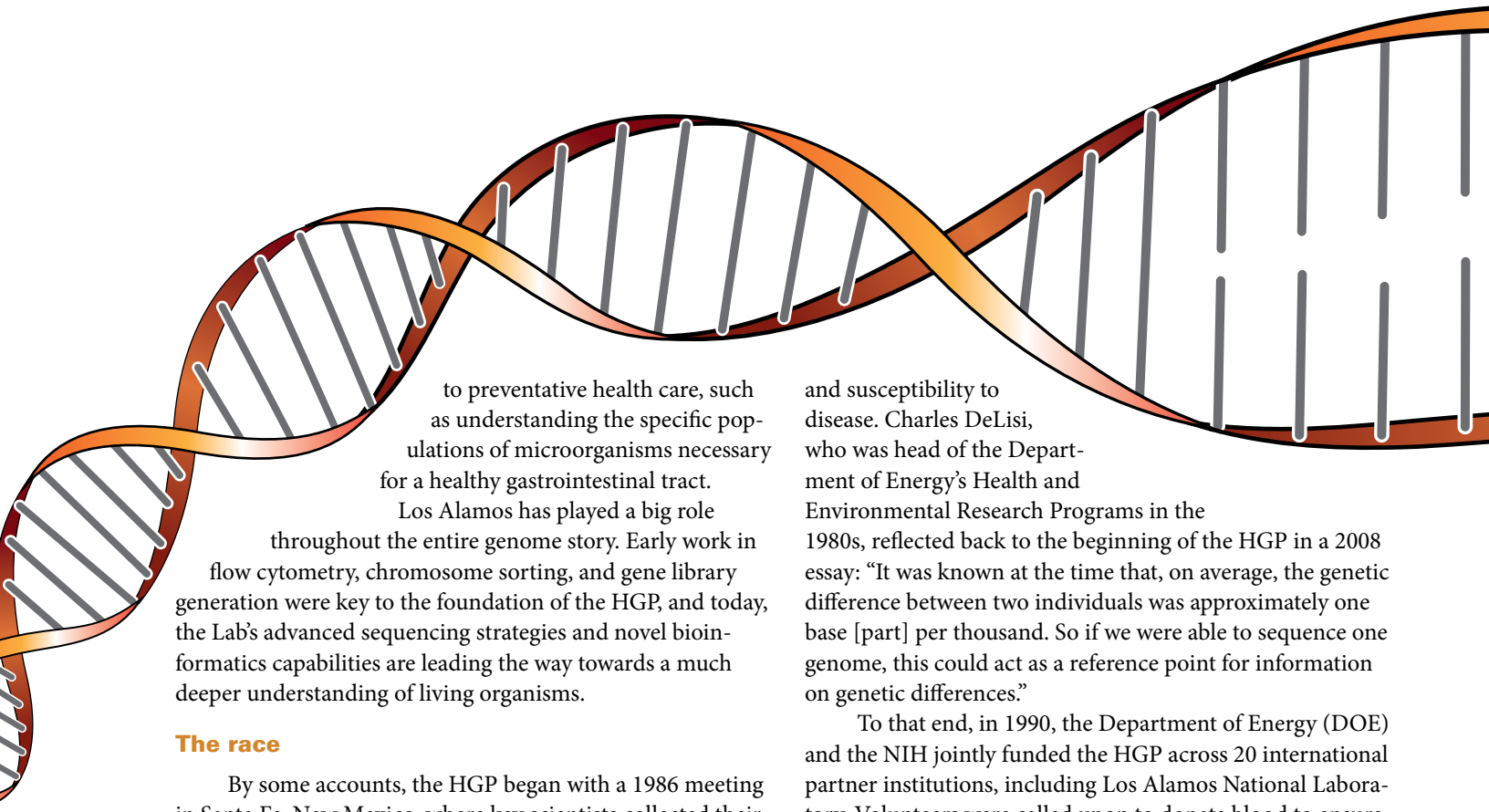
—Jay Schecker

Unraveling life four letters at a time

Ten years ago marked the official end of the Human Genome Project, but really it was just the beginning…

On April 24, 2003, the sequence of the human genetic code, or genome, was published, signifying the conclusion of the Human Genome Project (HGP). A major international effort, the project cost nearly $3 billion and was expected to take 15 years but finished two years early. It is a fascinating story of achievement and even a little drama.

The biggest achievement, however, may not reside in the actual map of the human genetic code, but in the genomics revolution that followed. In the 10 years since the map's completion, genetic sequencing has advanced at a phenomenal pace, revolutionizing the reach of biological research. The technology has become exponentially faster and orders of magnitude less expensive, and advances in bioinformatics are making it possible to thoroughly understand and analyze the mountains of genetic data now available. On the horizon are endless possibilities—from personalized medicine, such as cancer treatment tailored to the distinctive genetic fingerprint of each patient's tumor,



to preventative health care, such as understanding the specific populations of microorganisms necessary for a healthy gastrointestinal tract.

Los Alamos has played a big role throughout the entire genome story. Early work in flow cytometry, chromosome sorting, and gene library generation were key to the foundation of the HGP, and today, the Lab's advanced sequencing strategies and novel bioinformatics capabilities are leading the way towards a much deeper understanding of living organisms.

The race

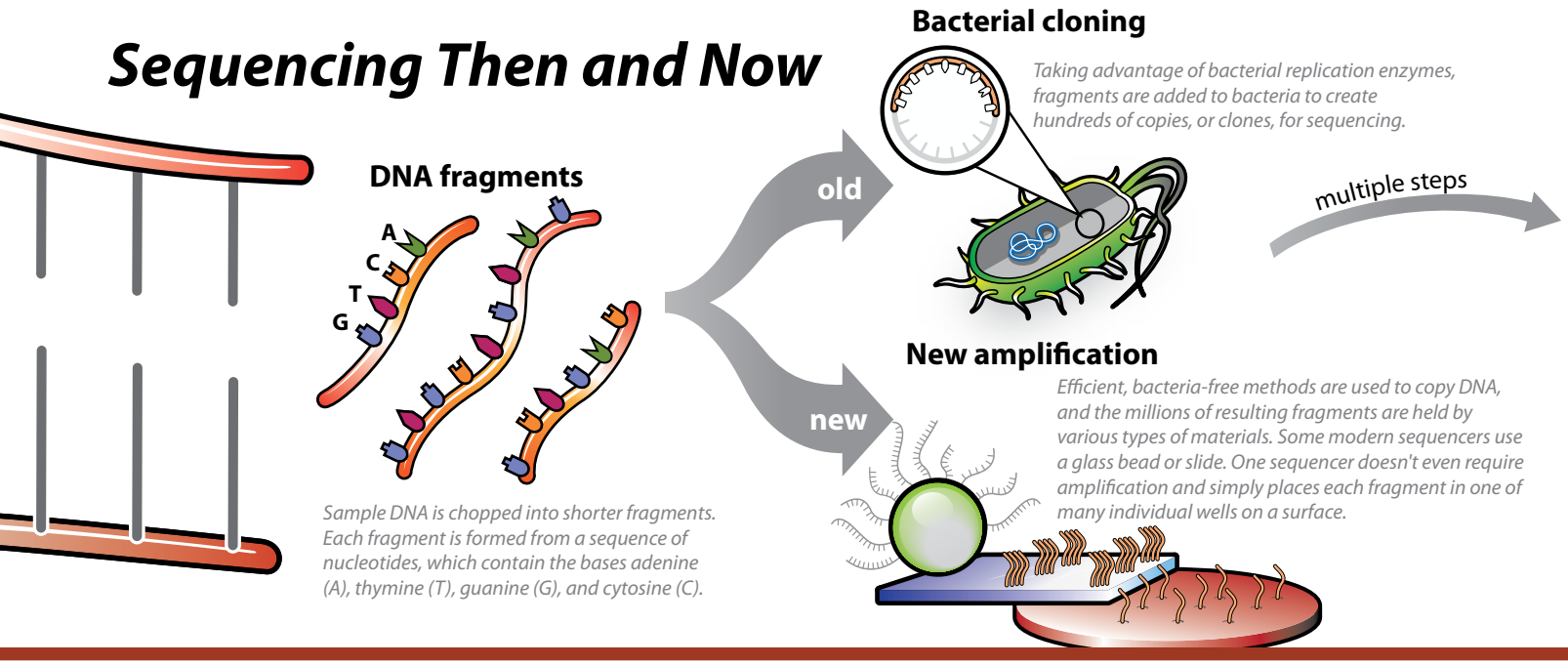
By some accounts, the HGP began with a 1986 meeting in Santa Fe, New Mexico, where key scientists collected their thoughts about why the government should fund the monumental endeavor to map and sequence all the DNA it takes to make a human. Scientists in Los Alamos had a long history of experience in this area, beginning with studying the mutagenetic effects of radiation on cells and leading to the National Laboratory Gene Library Project, in which they isolated, cloned, and packaged chromosomes into libraries for use by researchers worldwide. They accomplished this using flow cytometers, which were invented at the Laboratory [*see Los Alamos Firsts on the inside front cover of this issue of* 1663]. Also foundational to the HGP was Los Alamos's development of GenBank, now managed by the National Institutes of Health (NIH), a public database for genetic sequences.

"Leading up to the HGP, it was a very exciting time here at Los Alamos because we were doing stuff that very few other places in the world were doing," says Los Alamos biologist Jon Longmire.

At that time (the 1980s), a fair amount was already known about human genetics from various experiments and animal studies —multiple gene mutations had been linked to disease and localized to a particular chromosome. However, scientists knew there was much more to be learned about the mechanisms of inheritance, the functions of the human body, and susceptibility to disease. Charles DeLisi, who was head of the Department of Energy's Health and Environmental Research Programs in the 1980s, reflected back to the beginning of the HGP in a 2008 essay: "It was known at the time that, on average, the genetic difference between two individuals was approximately one base [part] per thousand. So if we were able to sequence one genome, this could act as a reference point for information on genetic differences."

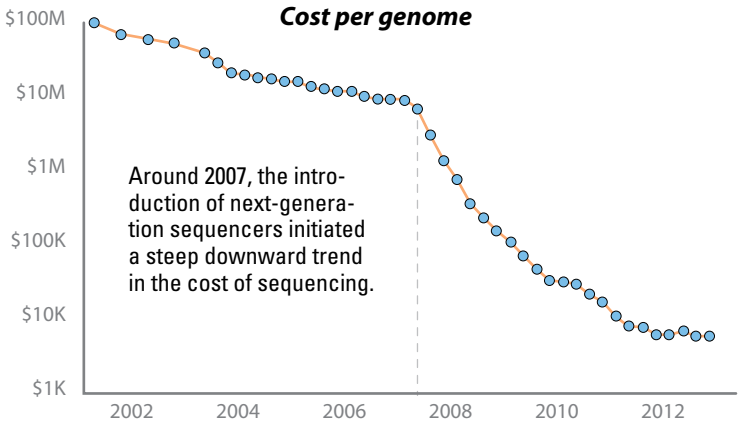
To that end, in 1990, the Department of Energy (DOE) and the NIH jointly funded the HGP across 20 international partner institutions, including Los Alamos National Laboratory. Volunteers were called upon to donate blood to ensure a diverse pool of starting material, the logic being that since human genomes are mostly identical, chromosome samples from a handful of individuals would give researchers a fairly accurate sense of an average person. Los Alamos's role was to create chromosome libraries from the donated material, with which each partner lab would map and sequence its assigned chromosome; Los Alamos would investigate chromosome 16 and part of chromosome 5. The HGP was expected to take 15 years using a laborious, incremental process that involved creating a physical map of each chromosome and sequencing and analyzing small sections of it in their correct order.

In 1998, the effort faced a challenger. A scientist named Craig Venter, who had previously worked with the NIH, created a company called Celera and proposed to complete the task in just three years using an alternative strategy that he had been using for microbial genomes. His approach broke up the entire genome (all chromosomes) into random small pieces, sequenced them, and then relied heavily on new computer algorithms to put the pieces back together in the correct order—a process known as whole-genome shotgun sequencing. Venter's method was far ahead of its time, and few believed it would work for such a large genome. However, it is now the basis of all modern sequencing.



Both public and private efforts published draft genomes in the top two scientific journals during the same week in 2001. The "race," effectively a tie, had been complicated by proprietary concerns, namely that the public effort was openly publishing its data while Celera was not, and questions abounded about patenting genes. The Clinton administration got involved and in March 2000 announced that the genome could not be patented. The "final" high-quality sequence was published in 2003, but in reality it remains a work in progress as many details are still being resolved.

By the end of the project, two breakthroughs had been made. Whole-genome shotgun sequencing had been proven successful, and a high-quality map of the human genome had been made. Initial analysis showed that the genome included about 30,000 genes widely distributed among many repetitive regions. Also found were many clues about recombinations and modifications that contribute to diversity. However, it was clear there would be much more to learn, and scientists are still investigating the complex roles of non-coding regions (sections of DNA that do not code for proteins as sections containing genes do).

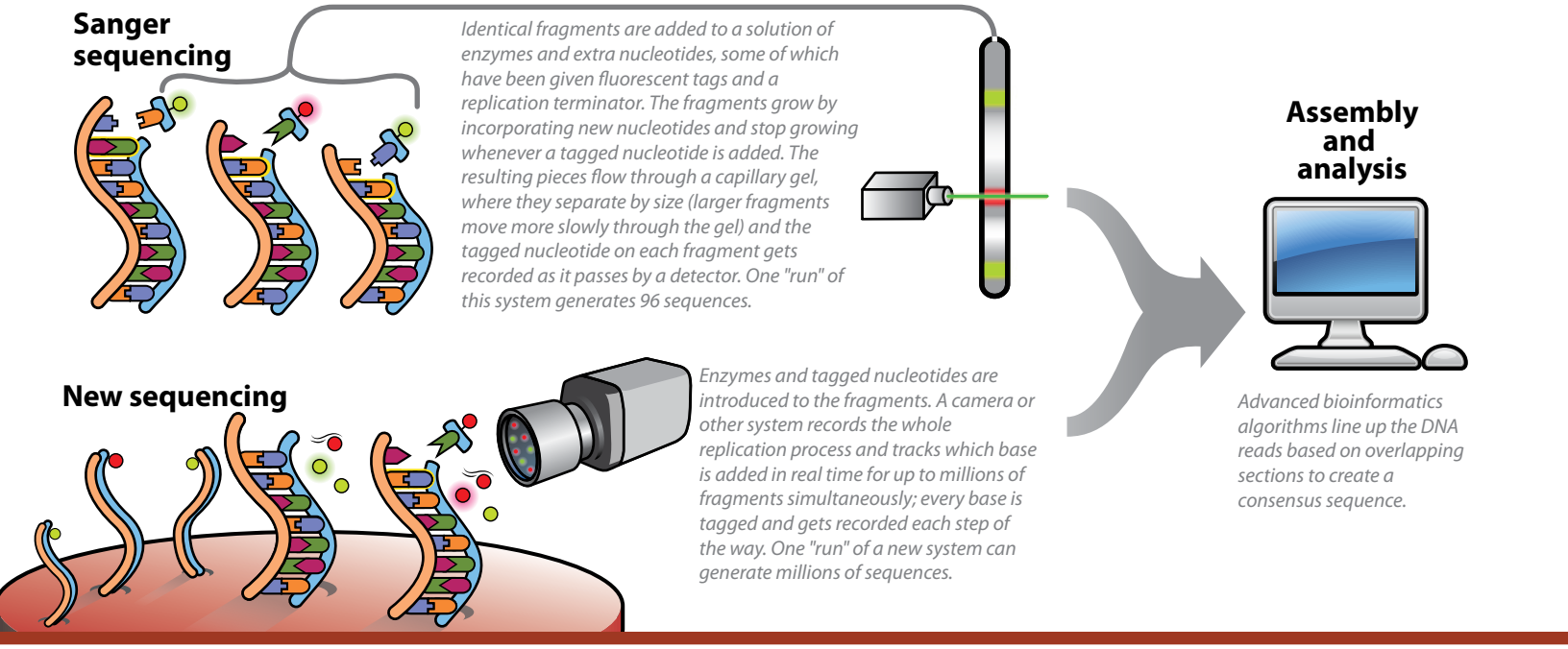


GATTACA, literally

In the 10 years since the publication of the human genome, major research institutions, such as DOE- and NIH-funded labs, have worked to understand how sequence data translates into an organism's characteristics and functions, while industry revolutionized the technology.

The now familiar "twisted ladder" structure of DNA was discovered in 1953. Its "rails" are made of alternating phosphate and sugar groups, and its "rungs" consist of nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The bases have a very specific binding protocol: A only binds to T, and C only binds to G. With only four bases in this code, the order, or sequence, of the bases is critical to understanding the meaning of the DNA. Methods of determining the sequence of bases began to arrive in the 1970s and the most well-known method, "Sanger sequencing"—developed by Frederick Sanger in 1977—was the primary technology used for the HGP and other projects for more than 20 years.

Most sequencing methods rely on the idea of mimicking DNA replication. During growth, cells multiply in number and must replicate their DNA as part of the process. When DNA is replicated, the ladder splits open into two template strands, and specialized enzymes called polymerases build new strands on the templates using free-floating subunits called nucleotides, each of which is a base plus phosphate and sugar—half a rung and a segment of rail. Laboratory sequencing methods strive to mimic this process and



identify which base is added when the new strand is built; the template strand is then inferred because of the binding protocol. For example, if a T is added, then the template must have an A at that location. However, spying on the activity of polymerases is difficult, so various sequencing strategies have been developed to make the task more tractable.

The first step in sequencing is to isolate and then shear the DNA into smaller, more manageable fragments. Next, most methods involve amplification, or making copies of the fragments. The identical fragments are sequenced and the results integrated to arrive at a reliable consensus sequence.

For traditional Sanger sequencing, the copied DNA fragments are exposed to polymerase enzymes, nucleotides, and a small number of tagged nucleotides that stop the replication process. (Early sequencing methods used radioactive tags, but in the 1980s, fluorescent tags were introduced.) The result is a large collection of pieces of identical DNA of different lengths, where each piece ends with a tagged nucleotide. The pieces then flow through a gel designed to separate them by size (larger fragments move more slowly through the gel), and the fluorescent tags get recorded as they pass by a laser detector. Together, the fragment sizes (organized by the gel) and nucleotide tags spell out the genetic sequence—fourth position T, fifth position C, and so on.

Modern, next-generation sequencing uses various approaches to parallelize the procedure. The result is a much cheaper process. Instead of bacterial cloning, new sequencers use more efficient methods of amplification, and some don't even need to amplify fragments at all. New sequencers use multiple strategies for tagging and detecting nucleotides and, instead of identifying each labeled nucleotide one at a time as fragments pass through a gel, use a camera or other device to chronicle which bases are added in real time for hundreds of thousands to millions of fragments simultaneously.

Once the fragments are sequenced, the resulting reads have to be assembled correctly in order to reconstitute the original genomic sequence. This is done by sorting through all the reads to find sections that can overlap. Once they are all lined up, the consensus sequence can be created. In the HGP, the chromosome maps helped with this process because researchers knew the physical location of a fragment before they sequenced it. But once shotgun sequencing became the dominant methodology, scientists began to rely completely on computers, using specific biological algorithms called bioinformatics (now a growing field in itself), to do it all.

"The birth of genome bioinformatics was because of both higher throughput and shotgun sequencing. We used to have to assemble hundreds of reads and now it's millions or billions," says Patrick Chain, a bioinformaticist at Los Alamos. Most next-generation sequencers use short reads, increasing throughput but requiring more bioinformatics to assemble. One notable exception uses single fragments (no amplification) and produces very long reads but, lacking redundancy, is more prone to errors.

"Reads," or sections of DNA produced by sequencing, are lined up according to their overlapping sections to create a consensus sequence of the entire genome.

Los Alamos genome scientists have become experts at creating complete, high-quality genome assemblies by combining the results of multiple sequencing platforms. Lab scientists also analyze these sequences and investigate the function of their genes for a diverse set of research problems, such as identifying the culprit in a disease outbreak or understanding the subtle differences between species.

Three billion's company

After the human genome was published, the NIH took on the challenge of deciphering the human data into more meaningful information, and the DOE, including its national laboratories like Los Alamos, dropped out of human genetics altogether. The DOE had created the Joint Genome Institute (JGI) in 1997 to unite the sequencing capabilities at DOE labs (Los Alamos, Lawrence Berkeley, and Lawrence Livermore), and in 2000 the JGI began to focus on sequencing microbes as related to DOE mission areas in carbon cycling, clean-energy generation, and environmental characterization and cleanup. This translated into years of work sequencing hundreds of microbes, which, when added to the public GenBank database, contributed to numerous comparative analyses.

One of the reasons microorganisms are interesting and valuable to study is that they live in a wide range of environments and are usually found in large, interdependent communities—interdependent upon each other and their surroundings. A huge variety of microbes in the soil, for example, can work in concert to degrade organic waste. Microbes are also important partners for other living organisms because of their ability to carry out functions to benefit their host environment, as is the case for the microbial community within the human gut that enables healthy digestion. In fact, microbes do quite a lot for humans, and there are about 10 times more bacterial cells on and in the human body than actual human cells.

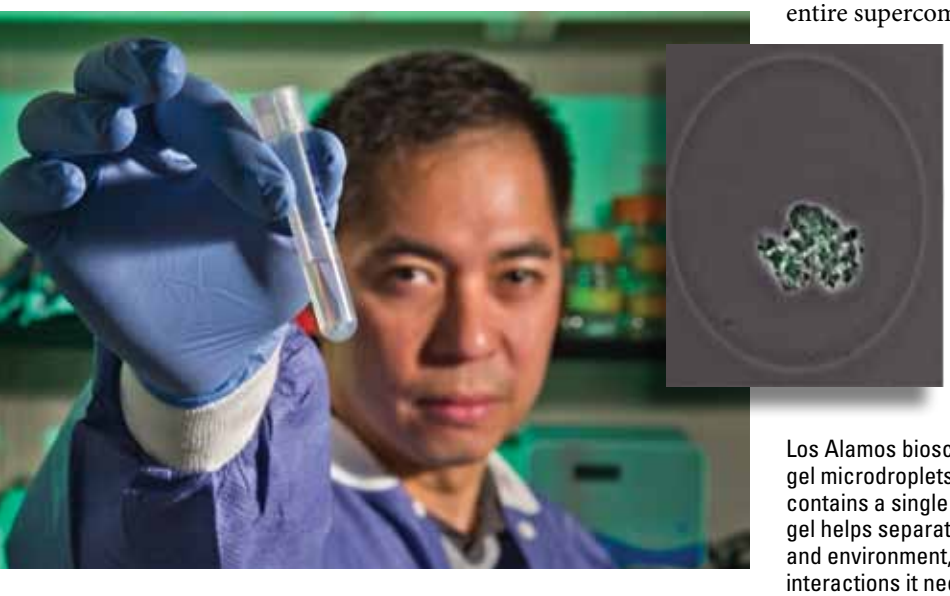
Although this interdependency makes them interesting, it also makes them very difficult to study because many rely on one another and their environments in order to grow. For that reason, isolating any specific type of microbe in a lab for sequencing can be difficult or impossible. The approach used to study these complex communities of microbes is called metagenomics, named for the idea of grabbing all the DNA in the community and treating is as one large *meta*genome. Metagenomics has become much more practical with modern sequencing.

"Now we can query more complex systems," says David Bruce, a manager for the genome group at Los Alamos. "Sequencing is no longer the rate-limiting step; now it's analysis."

In metagenomics, the challenge is to reassemble the billions of pieces that come from a wide variety of different organisms. Here, if some pieces reveal themselves to be from a known organism, scientists can assemble that organism's genome, but another tactic is to screen the pooled data for clues about what types of organisms are present. This is often achieved by organizing genes by certain known functions, rather than by organism. In addition to their arsenal of assemblers and other bioinformatics tools, Los Alamos scientists have designed two unique programs to tackle this challenge of classifying metagenomic data: Sequedex and GOTTCHA (Genomic Origins Through Taxonomic CHAllenge).

Both programs operate on the premise that specific kinds of signatures can help narrow the search criteria when comparing to known sequence databases. Sequedex focuses on amino acid signatures that are conserved—that is, common to two or more organisms or species. GOTTCHA focuses on nucleic acid (DNA or RNA) signatures that are unique to a genome or to a taxonomic group and tosses out all redundant genomic data. Both programs greatly reduce the amount of data that needs to be searched for matches, thus speeding analysis to the point that it can be carried out on a powerful laptop instead of an entire supercomputer.

Most sequencing requires a fair amount of DNA, so microbes have to be cultured and coaxed to replicate into a colony containing billions of cells. But since not all microbes can grow by traditional cultivation methods, researchers have developed new, potentially culture-independent techniques to isolate single cells from a mixed culture for sequencing. This approach, called single-cell genomics, is also promising for metagenomic samples—when a researcher may want



Los Alamos bioscientist Armand Dichosa holds up a vial of gel microdroplets. (Inset) Each tiny microdroplet (large circle) contains a single microorganism (green-gray shape). The gel helps separate the microorganism from its community and environment, while not completely isolating it from the interactions it needs to survive and grow.

Better Living Through Biology

CREDIT: ROCKY MOUNTAIN LABORATORIES, NIAID, NIH

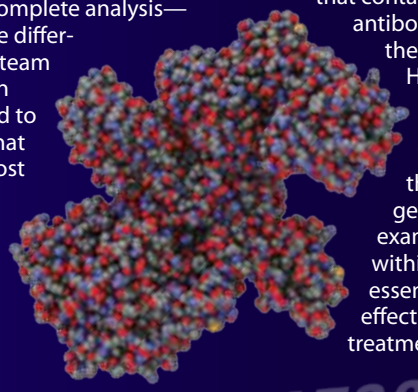
Identifying the cause of an outbreak

During the 2011 *E. coli* outbreak in Germany, Los Alamos scientists provided data analysis to prove the origin of the bacterial strain and why it was so virulent. The strain of *E. coli* carried a toxin normally found in Shigella bacteria and looked similar to one from a 2009 *E. coli* outbreak in the Republic of Georgia. By doing a complete analysis—comparing base-by-base differences—the Los Alamos team showed that the German strain was indeed related to the Georgian one, but that the German strain had lost the genes for one type of Shigella toxin and picked the genes for another.

Bad botox

Los Alamos scientists have characterized an extensive collection of the spore-forming bacteria *Clostridium botulinum*. This information assists public health officials in determining the specific type of botulinum neurotoxin that causes botulism in a patient.

"About 100 cases of infant botulism occur within the U.S. each year, and infants recover when provided a pharmaceutical product that contains a mixture of antibodies that neutralize the toxin," explains Karen Hill, the Los Alamos biologist who leads the project. "Identifying endemic strains that recur in certain geographic areas and examining the variation within the toxin types is essential for providing effective therapeutic treatment."

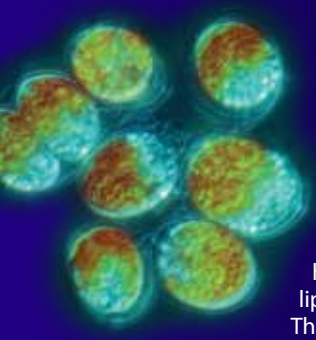


Soil secrets

Using metagenomics, Los Alamos scientists have led an extensive effort to study microorganism populations in various types of soil—from forests and arid lands to arctic permafrost. Many soil microbes are responsible for the degradation of organic matter (dead plants and animals), which essentially takes carbon out of the ground and puts it back in the atmosphere, a major part of the carbon cycle. The research team is working to determine the impact of climate change factors on these soil carbon cycling communities in the context of other ecosystem factors, such as soil and plant type, or regional climate.

DNA Genomics

ATCGGATCTAGGGATCCCATGCTAGCTAGCTCGATCGA
TAGCCTAGATCCCTAGGGTACGATCGATCGAGCTAGCTACGC



Biofuels production

Los Alamos microbiologists have been taking advantage of transcriptomics (sequenced RNA, indicating DNA that is currently in use) for their research towards creating biofuels from algae. By examining RNA transcripts from various strains of algae, at various time points, and under various conditions, they hope to identify genes involved in biomass or lipid production. (Lipids are used to make fuel.) They can then attempt to enhance growth and lipid production by adding these genes in the parts of the genome they know will be transcribed more often than others.

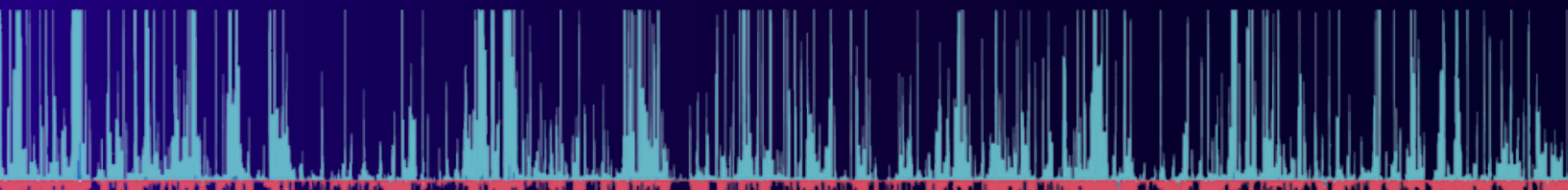
Personalized medicine

The Los Alamos genome team is collaborating with the University of California, Davis, to understand the molecular mechanisms that cause cancer cells to become resistant to chemotherapy drugs. Cancer is defined as the uncontrolled growth of abnormal cells, which is often caused by genes that are not being regulated properly. Since most regulatory molecules are coded by RNA, sequencing and quantifying the RNA of a tumor can help researchers understand what has gone awry in those particular cells. This process also narrows down the list of genes that might be responsible for chemotherapy resistance—expediting further experimentation.

RNA Transcriptomics

Examining infection

Transcriptomics can be used to examine interactions between organisms, such as when an organism has been invaded by a pathogen. The infection changes which genes are turned on or off by both participating organisms, and learning about these regulatory mechanisms can help scientists identify new targets for drugs or vaccines. To this end, Los Alamos genome scientists are examining the transcriptome of the bacterium *Yersinia pestis* (which causes plague) while it tries to evade destruction by a macrophage (an immune-system cell that engulfs invaders). The graphic below shows expression data from *Y. pestis* during this interaction; the red and blue spikes show the quantity of specific genes being expressed, giving an inside peek at the pathogen's defense mechanism.

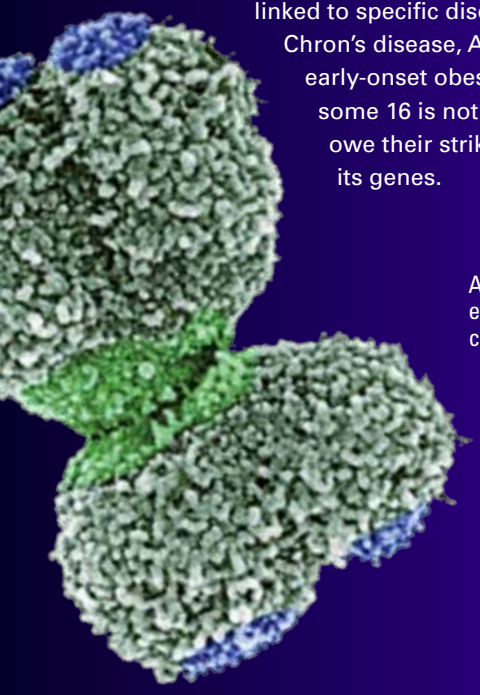


Chromosome 16,
Where Are You Now?

A major responsibility of the Laboratory during the HGP was to sequence the 90.4 million base pairs of chromosome 16 and to construct a detailed, high-resolution map showing the location of each of its approximately 1300 genes. The task was especially challenging because the chromosome has a larger-than-average proportion of duplications—long sections of DNA that are repeated on other parts of that chromosome or others.

Los Alamos biologist Norman Doggett, who was one of the principal investigators mapping chromosome 16, published a paper in 2006 (after the HGP had ended) about one of its duplications in particular, encompassing the HYDIN gene, which his team found to also exist on chromosome 1 (where it was named HYDIN2). It is one of the largest duplications across chromosomes in humans. Since the discovery of the HYDIN duplication, the HYDIN2 gene has been found to be associated with abnormal growth of the brain, causing developmental and behavioral problems, and the original HYDIN gene has been found to be responsible for primary ciliary dyskinesia, a rare genetic disorder that prevents the lining of the repiratory tract from removing mucus, exacerbating respiratory infections and conditions.

During the course of mapping chromosome 16, Los Alamos scientists contributed to the identification of several other important genes on the chromosome, including genes responsible for autosomal dominant polycystic kidney disease, Batten disease, and familial Mediterranean fever, to name a few. Today, more than 180 genes on chromosome 16 have been linked to specific disorders such as Chron's disease, Autism, and severe early-onset obesity. But chromosome 16 is not all bad; redheads owe their striking locks to one of its genes.



A colored scanning electron micrograph of chromosome 16.

to assemble an individual organism's entire genome even though the organism can't be cultured in isolation.

However, reliably assembling the complete genome from a single cell remains elusive; more DNA is needed. To this end, new Los Alamos technology allows the organism to be partially isolated in a droplet of gel for genomic study. The gel is porous enough for cell-signaling molecules (analogous to humans hormones) to flow in and out, such that the microbe is still able to communicate with other microbes and obtain nutrients from its environment. The captured cell grows into a microcolony of up to hundreds of identical cells while the gel keeps the microcolony sufficiently packaged for researchers to isolate the cells and assemble complete (or nearly complete) genomes in a high-throughput fashion.

DNA in action

Although DNA encodes the instructions for an organism's full potential, not all of the DNA is transcribed, or copied into RNA, so that it can be used. Liver enzymes, for example, are encoded by DNA in all human cells but are only transcribed in the liver. Therefore, the transcribed RNA tells researchers about the part of the genome that a cell is currently using, known as its transcriptome. The RNA is assembled with DNA as a template and contains all the information needed for subsequent translation into a protein.

By sequencing the RNA transcripts, researchers can find out more information about what a cell is doing and when, instead of just what it has instructions to do. Teasing out the tiny RNA molecules from a mass of genomic data has always been tricky, but the high throughput of some next-generation technologies, coupled with new bioinformatics and statistics, is making it easier. Current transcriptomics projects at Los Alamos include analyzing algae strains for enhanced production of biofuels, studying the molecular mechanisms of how cancer cells become resistant to chemotherapy drugs, and understanding the complex interactions between pathogens and their host organisms.

It's difficult to overstate the opportunities to advance biological research enabled by the explosion of technology since the official end of the HGP. Just ask Los Alamos genome scientist Momchilo Vuyisich, who has been studying the transcriptome from *Yersinia pestis* (the bacterium that causes plague) as it tries to outwit an immune-system cell. He explains that being able to examine the changes in gene expression during this interaction is a huge breakthrough. "If you showed this *Yersinia pestis* data to someone 10 years ago," he says, "they would think you were in a fantasy world."

LDRD

—Rebecca E. McDonald

PHASE

A NEW STATE OF MATTER

Why the cuprates? *That is the question.*

Cuprates, physicists would say, are certain ceramic metals made from copper, oxygen, and a variety of other elements. Like many metals, they are superconductors at temperatures close to absolute zero, but unlike *any other* metal, they remain superconducting at temperatures two to three times higher than their closest peer. The same physicists are baffled as to what it is about their atomic structure, their composition, or whatever else that enables them to create and sustain this high-temperature superconducting state?

Then there are the material phases. The elements making up a native cuprate are in certain proportions, which can be changed slightly by mixing in more of a particular element (doping). Doping often changes the electromagnetic properties of the metal, and depending on the degree of doping and the temperature, a cuprate will be in one of five material phases, or physical states: either a phase that exhibits an exotic type of magnetism (antiferromagnetism), the wonderfully bizarre high-temperature superconducting phase, an ordinary metal phase, a poorly understood metallic phase known as the pseudogap, or an equally poorly understood and strikingly weird metallic phase simply called a strange metal. These phases aren't unique; they are displayed by other (unconventional) metals, yet none of those materials are high-temperature superconductors. So why? Why the cuprates?

After 27 years of intense research, and more than 100,000 published research papers, condensed-matter physicists still can't answer that question. Neither can Los Alamos physicists Arkady Shekhter and Brad Ramshaw, but after conducting some pretty slick experiments this summer, the two scientists resolved a different longstanding question and gave theorists something solid to pin their hypotheses to. They showed conclusively that the pseudogap region of the cuprate phase diagram is a distinct phase of matter.

"For years, physicists thought that the pseudogap wasn't a distinct phase, just a region where the physical properties of the strange metal continued to evolve as the material cooled," says Shekhter. "The experimental evidence for a distinct phase was not compelling—the measured data had large uncertainties. With our measurement, there is no room to wiggle. Our data cannot be ignored."

Proof that the pseudogap was a fifth thermodynamic phase had immediate ramifications. It strongly suggested that the key to understanding high-temperature superconductivity would be found within the physics of the strange metallic state.

The backstory

In 1986, superconductors got high for the first time. Superconductors are materials that when cooled to extraordinarily low temperatures enter into a magical state in which electrons flow effortlessly through the material without loss of energy. Electric motors stay cool, power lines transport current with 100% efficiency, and electromagnets produce super-strong magnetic fields that can levitate anything from frogs to passenger trains. When electrical engineers daydream, superconducting circuits dance in their heads.

The problem was that a superconductor only superconducts when cooled below some critical temperature, denoted by T_c. The highest known T_c was about 23 degrees above absolute zero (23 kelvins, or 23 K). It required an unwieldy, expensive coolant—liquid helium—to reach that state of freeze. Consequently, superconductivity was a remarkable but seldom exploited phenomenon.

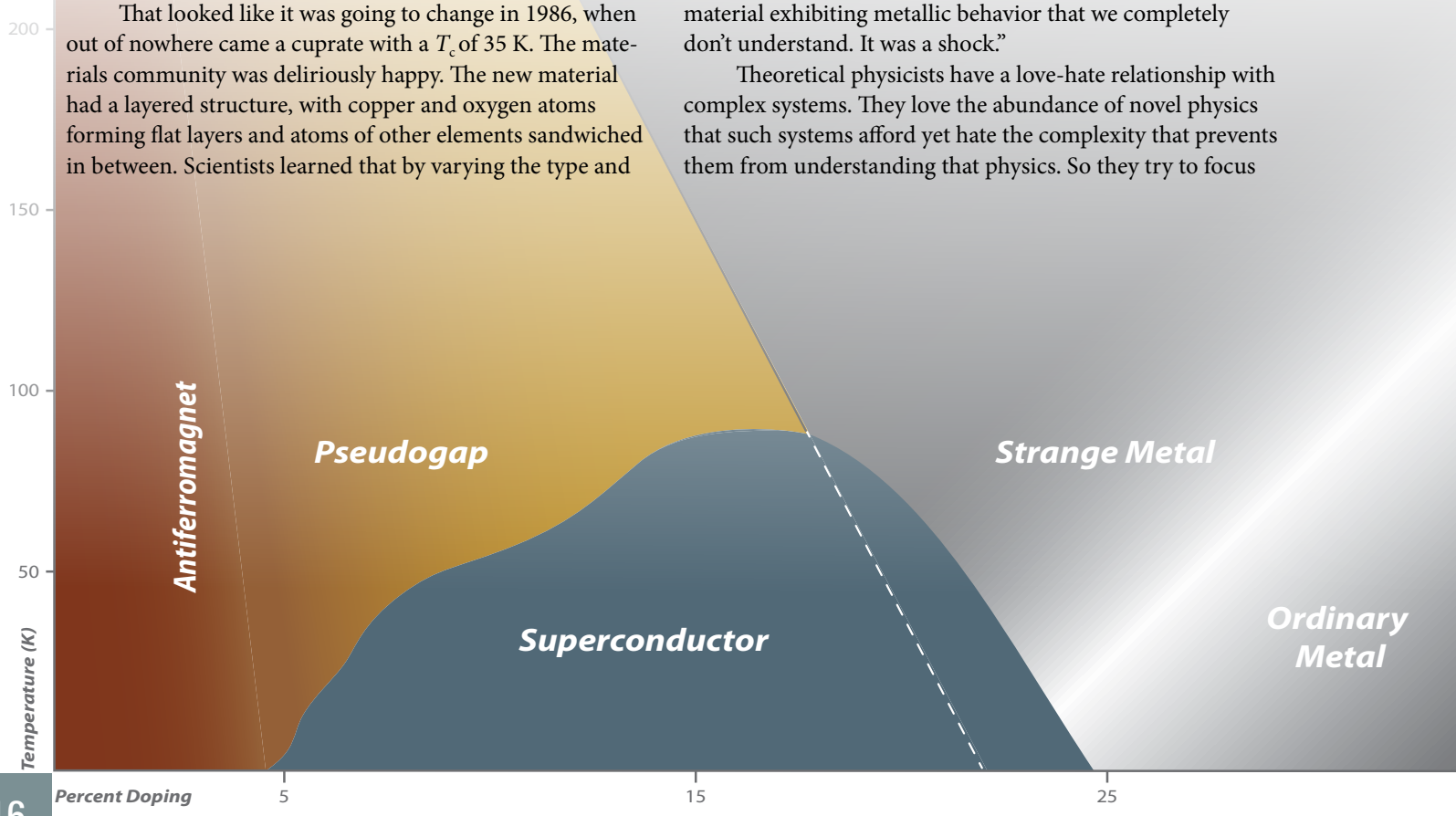
That looked like it was going to change in 1986, when out of nowhere came a cuprate with a T_c of 35 K. The materials community was deliriously happy. The new material had a layered structure, with copper and oxygen atoms forming flat layers and atoms of other elements sandwiched in between. Scientists learned that by varying the type and proportions of the other elements, they could change the critical temperature. Within two years, critical temperatures had climbed to 138 K, and scientists (and venture capitalists) could just about taste the fruits that would be harvested from a 300 K, room-temperature superconductor.

But there was no fruit. Other than by applying absurdly high pressures (hundreds of thousands of atmospheres), researchers could not increase T_c beyond 138 K. Furthermore, the cuprates were bizarre, and their superconductivity different from that manifested by conventional superconductors. They couldn't carry as much current as was hoped and would cease to be superconducting in even modest magnetic fields.

Most startling was their behavior at temperatures above T_c. The first measurements of cuprate single crystals showed that while the materials were metals, they were strange metals. For example, ordinary metals have a finite electrical resistivity at high temperatures, because thermal energy makes their atoms jiggle in place, which impedes the flow of electrons. The jiggling eases as the temperature drops, and the resistivity drops as described by "Landau's Fermi-liquid theory," which successfully explains the low-temperature behavior of conventional metals. But in the strange metal phase, the cuprates' resistivity didn't depend on jiggling atoms at all. That identified the cuprates as members of a class of metals known as "non-Fermi-liquid" metals. As they are the only metals to exhibit high-temperature superconductivity, they're really in a class of their own.

"The cuprates were unique," says Shekhter. "At that time in the mid-1980s, everyone believed they knew everything there was to know about metals. And suddenly there's this material exhibiting metallic behavior that we completely don't understand. It was a shock."

Theoretical physicists have a love-hate relationship with complex systems. They love the abundance of novel physics that such systems afford yet hate the complexity that prevents them from understanding that physics. So they try to focus



Brad Ramshaw (left) and Arkady Shekhter.

on one property and build the simplest model that captures the phenomena. As the model matures and becomes robust, they try and expand it to include more physics in the hope of explaining more and more phenomena with it.

Over time, many physicists came to view the strange metal phase as the core feature that distinguished the cuprates from other materials. While technologists continued to roam the high-T_c landscape, hoping to find the path to a room-temperature superconductor, many other scientists ignored the high-T_c superconducting phase altogether as they tried to understand the strange metal state.

"Physicists were completely stymied by this weird, strange metal state," says Ramshaw. "They had no idea who the players were that were responsible for its properties. Were they electrons, quasiparticles, strange couplings between spins and the lattice? No one had a clue."

Not knowing what to make of the strange metal state simply prompted more theories about how to make it. One of the more promising suggested that this solid phase was quantum-critical—the material's bulk properties depended on the details of how its quantum states evolved in time. This is not true for most phases. In a non-quantum-critical phase, the material is equally likely to be found in any of its available quantum states, so the physical properties only depend on what those states are, but not on how the system gets from one quantum state to the other.

If a quantum-critical phase were part of the cuprate's story, it would mean that as the temperature changed, a cuprate would undergo a phase transition between quantum-critical and non-quantum-critical phases. That transition should be observable at a specific temperature as a sharp change, or discontinuity, in many material properties.

Depending on the temperature and its specific composition (doping), a cuprate will be in one of five material phases. Undoped, the material is an insulator and an antiferromagnet, but it becomes metallic and an increasingly better metal as doping increases. Above 5 percent doping, the cuprate will exhibit superconductivity up to the critical temperature, T_c. At higher temperatures, it will either be in the pseudogap or strange metal phase. Above about 25 percent doping, the material is an ordinary metal. Shekhter and Ramshaw proved that the pseudogap was a distinct material phase.

The stage was set for Shekhter and Ramshaw to make measurements of a cuprate property and look for a discontinuity as the material cooled down. For the experimental method, they turned to an old workhorse—resonant ultrasound spectroscopy (RUS).

A new phase

Resonant ultrasound spectroscopy has been used for decades to study how the elastic properties of a crystalline material evolve with temperature. Briefly, one sandwiches a single crystal between two ultrasound transducers, one to vibrate the crystal at a chosen frequency and the other to detect the crystal response. At a resonance, the crystal vibrates strongly. By slowly changing the vibration frequency, one can find and measure all of the crystal's resonance frequencies, from which one determines the material's elastic constants. Changes in the elastic properties are tied to changes in the thermodynamic state of the material, including those associated with a phase transition.

Mentored in RUS by Los Alamos Fellow Albert Migliori, who looms large in the development of the technique, Shekhter and Ramshaw set about adapting it to work at low temperatures with very pure but uncomfortably small crystals. Small crystals meant small signal. Even at resonance, their response would be completely overwhelmed by the noisy vibrations of the surrounding environment. Experimental success boiled down to finding structural materials that were strong enough to hold the transducer-crystal sandwich, yet soft enough to isolate the sandwich from the rest of the vibrating world.

The right material turned out to be balsa wood, the stiff-yet-lightweight staple of the model airplane industry. The physicists began taking data, using a revolutionary new data acquisition and analysis technique they had developed specifically for this measurement. In theory, it would afford them significantly better temperature and frequency resolution.

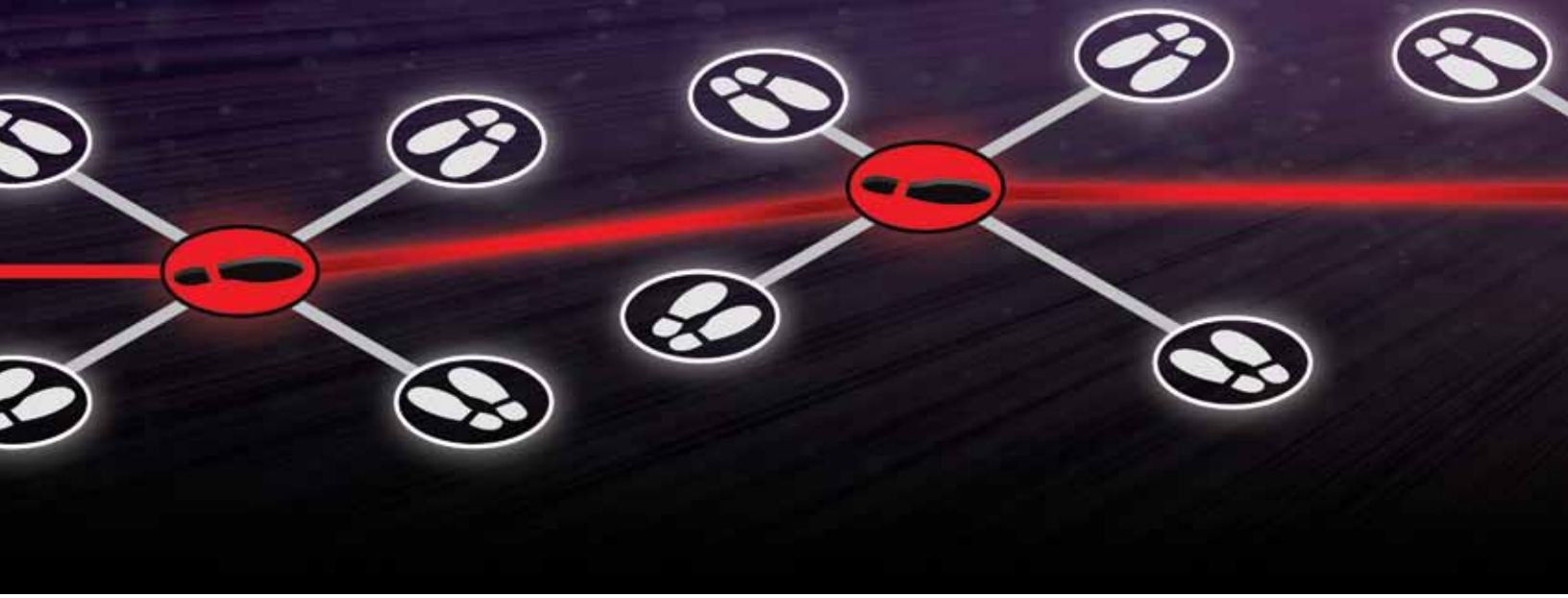
"It worked," said Shekhter. "Our error bars were plus or minus 3 K, a factor of 10 tighter than people were able to do using other techniques."

While few now doubt that the pseudogap is a separate phase, and while the evidence for quantum-critical behavior is stronger, there's still no clear answer to what makes the cuprates special. It may be a simple answer, like the energy levels of copper are closer than any other transition element to the energy levels of oxygen, so a copper oxide, as opposed to any other elemental pairing, is the only combination wherein the electrons can arrange themselves as needed to produce strange metal states or high-T_c superconductivity. Or it may be something else. The only thing definitive is the question, "Why the cuprates?" LDRD

-Jay Schecker

INTRUDER ALERT FOR THE CYBER WORLD

SOPHISTICATED HACKERS ARE GOING TO GET IN. THE TRICK IS TO FIND THEM ON THE



On January 7, 2013, Los Alamos National Laboratory was hacked by an unlikely adversary: itself.

"It's a case of good guys acting like bad guys to test our defenses," explains Josh Neil, a cyber security statistics expert at Los Alamos. "In this case, it was our own Engineering and Security Services group." Neil leads a project that develops software to protect against cyber attacks. The software, called PathScan, has been up and running at Los Alamos for about a year, ready and able to catch intruders. Indeed, it proved capable of spotting the friendly attack and continues to staff its cyber sentry post, ever watchful for unfriendly ones.

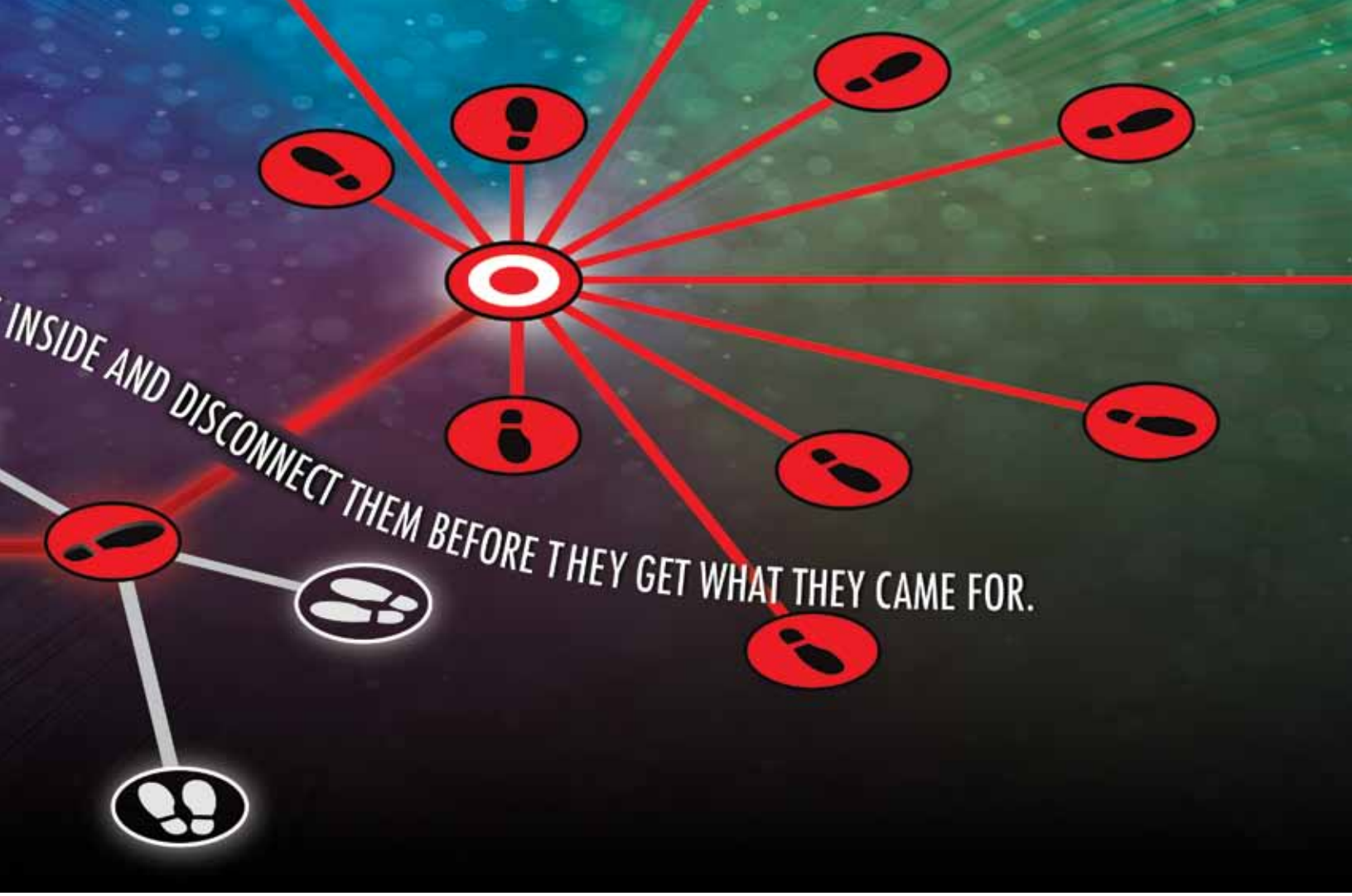
Can it do the same for other protected networks? All evidence points to yes. It has already run or is currently running in a trial phase on other government and industry networks, with another trial run about to begin.

"I think we've got a ground-breaking technology for network defense," Neil says. "This could really change the cyber security landscape for the better."

Byte background

Conventional cyber security software works by scanning network activity for particular packets of information, or byte strings, that correspond to previously reported cyber crimes or other malware activity. Because this approach matches current network byte strings to well-identified malicious byte strings, it is extremely accurate at diagnosing known attacks. In this sense, it is similar to antivirus software running on an individual computer. It affords reliable protection against a long list of established viruses despite its vulnerability to something new. But that's where the virus analogy ends.

"Sophisticated attacks generally are not virus-like," says Neil. "Exponentially exploding, automated intrusions may sound worrisome, but in practice, they're the easier ones to catch."



Even if the particular form of a virus-like attack on an individual computer has never been encountered before, its explosive nature still tends to give it away on the network scale: it radiates outward from one computer to many, forming a pattern known as a star. Such an attack can be detected rapidly, with safeguards automatically engaged upon detection. For example, the affected computers might have their network connections disabled while alert messages are dispatched to network administrators. Attacks that can be handled in this way aren't the ones that most concern Neil.

"The ones you really have to worry about are those being driven by a human being in real time," he says.

In contrast to a virus-like attack that spreads rapidly from computer to computer and therefore induces a fairly obvious abnormality in the network usage pattern, a single user can try to hide in the daily bustle of network activity. The Los Alamos unclassified computer network, for example, contains about 20,000 computers generating more than 500 million communication events (from one computer to another) every day—presenting an enormous background in which a hacker may effectively disappear.

The main virtue of PathScan is its ability to find hidden hackers by recognizing the subtle, small-scale network abnormalities that result from their intrusions. It does this as its name implies, by scanning for the paths taken by hackers as they move around within the target network.

Your enemies closer

Hackers typically want to steal digital information, including personal and proprietary information. In some cases they may be individuals who intend to use the stolen information themselves or sell it for profit. In other cases they may be well-funded nation-states that plan to use the stolen information for national gain.

The first line of defense against such cyber theft is a firewall. A firewall analyzes data packets entering a protected network and rejects those that pose a security concern, such as external login attempts without proper authentication. As a result, a hacker can't access computers beyond the firewall directly and must instead obtain some kind of insider access. A brute-force approach might be to physically break into a

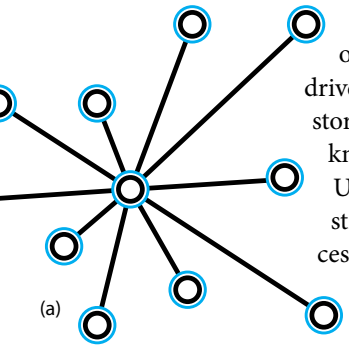
facility and stick a malware-containing USB drive into a computer. But a more common (and less risky) approach is phishing: sending an official-looking email to employees on the inside that encourages each employee to click on an external link or open an attached file. Either action delivers malware to the employee's computer. Because the employee accesses the link or the attachment deliberately—that is, the connection is initiated from within by an authorized user—the firewall may not prevent the subsequent malware download. This allows the external hacker to access the employee's computer.

The most prevalent way defend against phishing attacks (without disallowing all web activity originating from email clicks) is to train employees to recognize suspicious emails. But this defense is not perfect. "We can't rely on being able to stop every phishing attack from getting through while maintaining current network usability," says Neil, "and that means we have to be able to detect and stop an intruder inside the network."

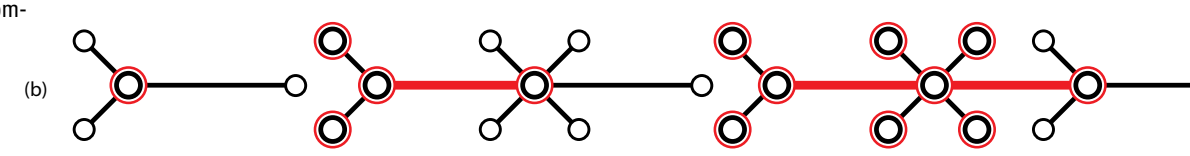
Crawling criminals

Fortunately, an employee who gets tricked by a phishing attack rarely has whatever the hacker is looking for right there on his or her computer. That means the hacker must hop from the hacked computer to other computers on the network in search of data worth stealing. This usually isn't quick or easy, because not every computer can access every other computer.

Once a hacker arrives at one computer, he or she has the ability to log in to all the other machines normally accessible from that computer because login credentials to those other machines are stored on its hard drive. (Passwords may not be directly stored, but encrypted versions of them, known as hashed passwords, are. Unfortunately, hashed passwords can still provide login access in a process known as "pass the hash," even if the hacker can't read the original passwords.)



Cyber attack patterns: (a) A star formation occurs when an attacker, whether human-driven or self-replicating like a computer virus, uses one hacked computer to reach out to many more, checking each for vulnerabilities, useful data, or login credentials to other computers. (b) A caterpillar is a human-controlled attack formation in which the hacker progresses from each computer to a handful of others. Moves that further the hacker's goals form the caterpillar's body, while backtracked dead-ends form the legs.



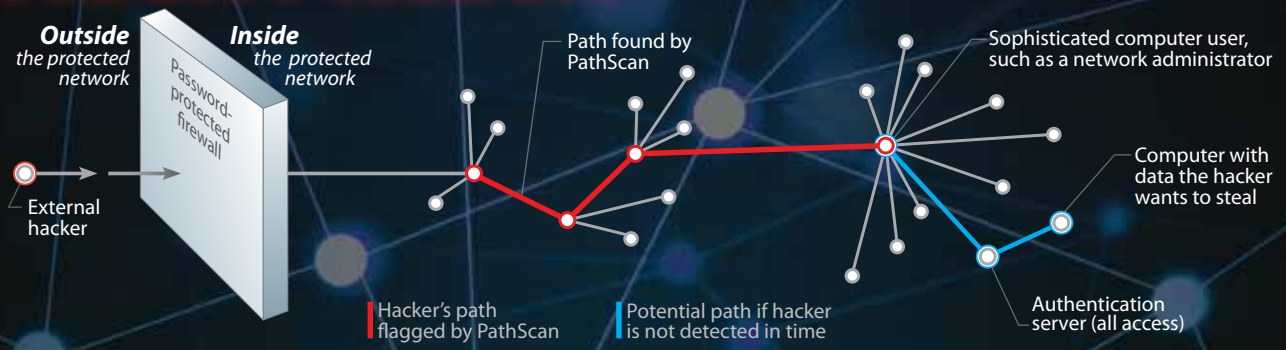
Typically, the credentials stored on one computer deliver access to many other machines, including email servers, network drives, shared printers, and even other office computers. These are not always useful to the hacker, however, partly because the login credentials obtained from one computer usually convey few access privileges to the files on another computer—effectively "look but don't touch"—and partly because the most sensitive machines require progressively higher-level credentials.

As a result, a hacker must hop from one machine to another in search of more privileged login credentials, advancing (and sometimes backtracking) across the network. This creates a tell-tale pattern of motion, hopping from one computer to several others, settling on the most promising one, and repeating the process from there. The pattern, when drawn on paper, appears as a line of "correct" hops with a bunch of "incorrect" hops extending off to the sides. The pattern is named for its resemblance to a caterpillar, with the body formed from the line of correct hops and the legs formed from all the others. PathScan searches for these caterpillars.

The process of navigating from one computer to the next is slow: the hacker has to examine each machine to see what information and login credentials it stores. And most computers on the network have little or no useful information to steal and no elevated login credentials beyond what the hacker already used to get that far. However, some users are more sophisticated in terms of their network use, such as system administrators who oversee all the computers in some large group within the organization. If an intruder successfully enters one of these sophisticated-user machines, he or she will find a set of login credentials with complete administrative access privileges to more valuable target servers. With these privileges, the hacker can view or copy anything and even install permanent programs like backdoors, which allow the hacker to secretly return to a machine with full adminstrative access at any time in the future.

There is one machine in particular that the intruder really wants to access, known as an authentication server. It holds and recognizes the login credentials for users all over the organization and is involved every time any computer on the network authenticates to any other. This machine is the hacker's ultimate all-access pass; if a hacker invades the authentication server with a valid administrator credential, every other computer becomes available, including those containing the data the hacker wants to steal. Therefore, the trick is to detect the caterpillar's motion and shut it down

Hack Attack



A successful cyber attack, usually attempted for the purpose of stealing proprietary data, requires several steps:

(1) The attacker must access a computer inside the protected network by establishing a connection across the organization's firewall. This requires password authentication, and if the hacker can't obtain valid login credentials, he or she may make a phishing attempt instead. The attacker contacts an employee of the organization by email and tricks the employee into clicking on something malicious—either an attached file or a link to a compromised website—which then allows access to that employee's computer.

(2) Since the information the attacker is after is unlikely to reside on the first computer reached, and because the original firewall crossing might have alerted computer security personnel to the intrusion, the attacker must hop to another computer using login credentials found on the original hacked computer. In this way, the intruder moves from one computer to another, continually searching for higher-level login credentials that convey elevated access privileges on protected systems and servers.

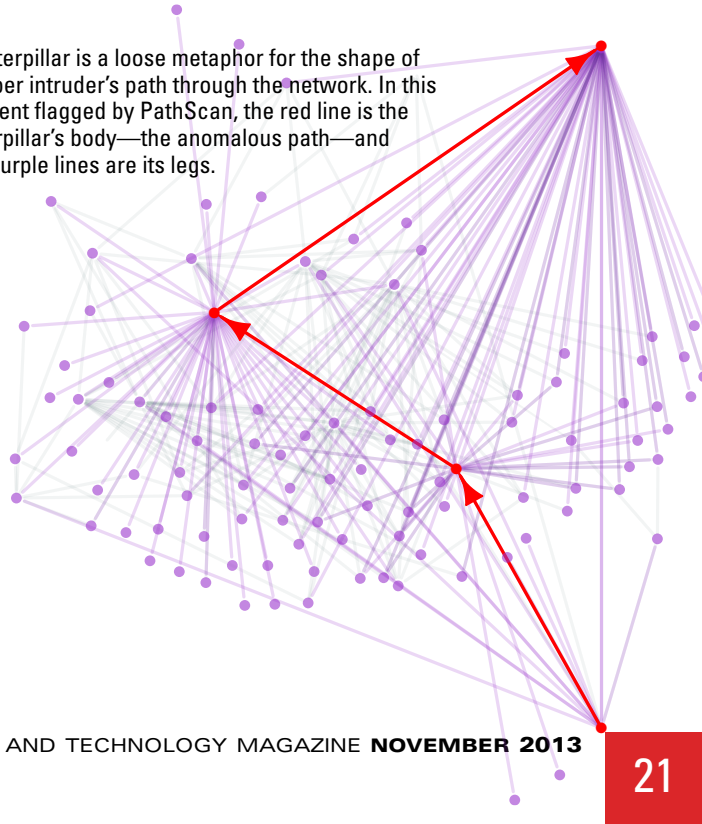
(3) Any combination of three consecutive hops considered out of the ordinary by the Los Alamos PathScan software (when compared to normal network activity) alerts computer security personnel to investigate. They determine the extent of the attack and how to respond. It may be necessary, for example, to disconnect part or all of the organization's network from the outside world, wipe certain computers clean, or replace certain logins. (The red line illustrates a detectable intruder path.)

(4) With each hop, the attacker tries to move toward a computer with greater access, such as a computer used by a network administrator or, better yet, the organization's authentication server, which stores login credentials to all other computers on the network. If the attacker is not detected before reaching these machines, then the computer containing the data he or she intends to steal will become accessible. (The blue line indicates this access sequence if the attack is not stopped earlier.)

before it reaches the authentication server. Otherwise, if that server is hacked, the entire network will have to be taken offline while every login in the whole organization is changed and every computer in the organization is analyzed for tampering and theft. It may even be necessary to wipe many computers completely, which would represent a substantial cost in terms of downtime and lost productivity, even if the data theft was ultimately unsuccessful.

Guarding the edges

Neil and his PathScan colleague Curtis Storlie are statisticians, and PathScan is primarily a probability and statistics analysis tool. It maintains a baseline statistical pattern for normal network behavior and identifies network communications that deviate from that pattern. What is the prob-



A caterpillar is a loose metaphor for the shape of a cyber intruder's path through the network. In this incident flagged by PathScan, the red line is the caterpillar's body—the anomalous path—and the purple lines are its legs.

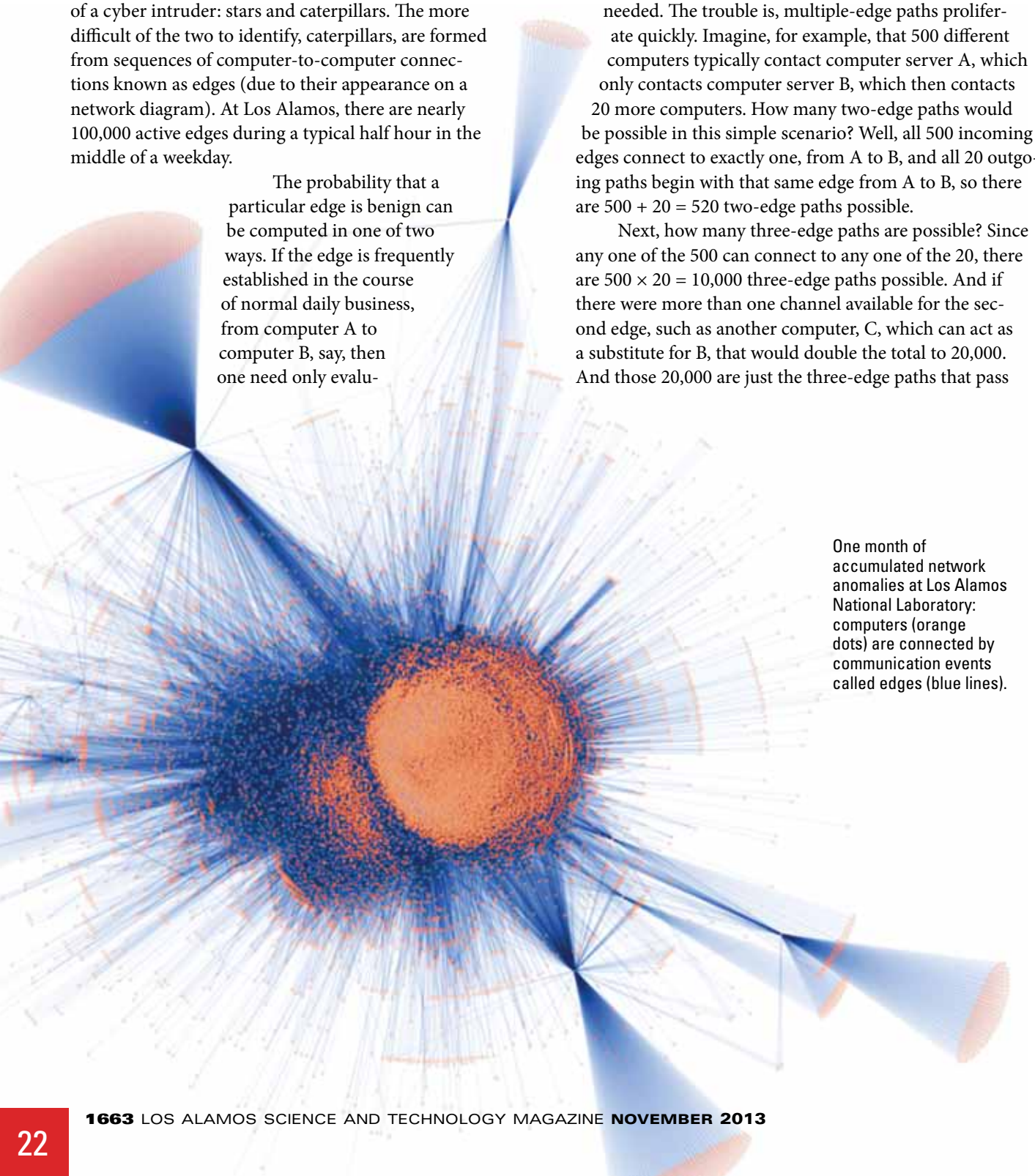
ability that computer A contacts computer B at a particular time on a Wednesday afternoon? What if it's the third time A has contacted B in the last hour? What if A just contacted C 10 minutes earlier, and D 25 minutes before that? And what if B contacts E after being contacted by A? Are these events normal or suspicious?

Also on the team are system architect Curtis Hash, large-scale data-management specialist Alexander Brugh, and cyber security expert Mike Fisk. Together, the team has designed the software to sift through an enormous volume of network data in real time, looking for the major footprints of a cyber intruder: stars and caterpillars. The more difficult of the two to identify, caterpillars, are formed from sequences of computer-to-computer connections known as edges (due to their appearance on a network diagram). At Los Alamos, there are nearly 100,000 active edges during a typical half hour in the middle of a weekday.

The probability that a particular edge is benign can be computed in one of two ways. If the edge is frequently established in the course of normal daily business, from computer A to computer B, say, then one need only evaluate the frequency. If that connection is typically established 50 times in a given time period, then how unusual would 75 be? A question like this can be answered probabilistically. Alternatively, if the edge is completely new because A has never contacted B before, then the probability that it's benign is modeled as a logistic regression, blending three distinct rates: the rate at which new edges appear on the network overall, the rate at which A initiates them, and the rate at which B receives them.

But individual, anomalous edges are usually insufficient to indicate an intruder; rather, multiple edges are needed. The trouble is, multiple-edge paths proliferate quickly. Imagine, for example, that 500 different computers typically contact computer server A, which only contacts computer server B, which then contacts 20 more computers. How many two-edge paths would be possible in this simple scenario? Well, all 500 incoming edges connect to exactly one, from A to B, and all 20 outgoing paths begin with that same edge from A to B, so there are 500 + 20 = 520 two-edge paths possible.

Next, how many three-edge paths are possible? Since any one of the 500 can connect to any one of the 20, there are 500 × 20 = 10,000 three-edge paths possible. And if there were more than one channel available for the second edge, such as another computer, C, which can act as a substitute for B, that would double the total to 20,000. And those 20,000 are just the three-edge paths that pass



One month of accumulated network anomalies at Los Alamos National Laboratory: computers (orange dots) are connected by communication events called edges (blue lines).

through computer A after the first edge. On the Los Alamos unclassified network, mid-day on a weekday, a typical 30-minute window contains about 300 million three-edge paths.

Based on their extensive cyber experience, Neil and his team made the decision to stop there rather than include four- and five-edge paths. Their reasoning? Not only does the increase in path length beyond three make the number of paths more computationally expensive, but it also requires that intruders make more moves to get caught and therefore misses those who achieve their objective in fewer moves.

With so many three-edge paths to examine, PathScan needs to be very discriminating about which ones it considers suspicious. It does this by computing a probability for each path it observes on the network: the probability that the path in question should emerge at that given time, assuming that no cyber attack is actually underway. That is to say, under normal business conditions, what is the probability of a particular path occurring? If the probability is too slim, PathScan generates an alarm, and cyber security personnel investigate the potential intrusion.

How slim is too slim? That's open to debate. If the threshold probability is set too low, then very few anomalous paths will be reported, and it may be possible for an intruder to slip by. If it is set too high, that will create additional workload for analysts chasing down false alarms. Additionally, the optimal sensitivity setting may differ from one organization to the next, with a national security laboratory like Los Alamos choosing a relatively more conservative threshold, calibrated against historical network data collected during the past 10 years. Within that data, PathScan identified several sophisticated attacks. Such sophisticated attacks do not come along often, and PathScan has demonstrated exceptional reliability at isolating these exceedingly rare events within the mountains of data processed.

Path forward

Pilot programs for PathScan have been or are being conducted at a number of organizations other than Los Alamos, including the U.S. military and the oil and gas industry. And through a partnership with the Department of Homeland Security's Transition to Practice Program, another pilot program is set to begin at the Department of Veterans Affairs in the coming months. Upon the successful conclusion of these pilot programs, the team intends to deploy PathScan more widely. But that won't be the end of the story. Cyber criminals present an ever-evolving threat, and cyber security efforts have to keep up. PathScan must continue to evolve as well, and Los Alamos remains the best place for that to happen.

"Los Alamos is a leader in network data collection," says Neil "I couldn't do this anywhere else. The data we track—bytes and packets sent and received, types of communications occurring—usually doesn't get comprehensively reported, even at very wealthy companies. The time has come for other enterprises to collect and analyze their network data more effectively, as we do."

To stay ahead of cyber crime, PathScan is learning to search for additional sorts of anomalous events captured by these network statistics. Imagine, for example, that the exact same number of bytes is sent from computer A to computer B, then C, then D. Even if the path ABCD doesn't register as anomalous, the constant byte count would. "If that transmission were legitimate, it would have gone straight from A to D," Neil says. His team is already testing prototype enhancements of this sort.

The next step will be training PathScan to figure out on its own what to look for, based on real-time network usage. Although anomalous three-edge paths are a strong indicator of malicious activity, under certain network conditions, other indicators might do even better. The team wants to create models of new and ingenious cyber attack methodologies and teach PathScan how to adapt and reprioritize its search objectives accordingly, on the fly. Neil is convinced this can be done and hopes that cyber security can outpace cyber crime as a result.

"Lately, you hear about successful cyber attacks on high-profile company networks, and you get the impression we're losing all these individual battles," he says. "But I think we can defend our networks and the intellectual gold they contain much better than we do now—and get some real cyber defense wins." LDRD

—Craig Tyler

The Los Alamos PathScan team, clockwise from bottom, is Josh Neil, Curtis Storlie, Curtis Hash, Mike Fisk, and Alexander Brugh.

spotlights

Moulin Bleu

Most of the time, new results in climate science seem to be bad news. Sea levels are rising faster than previously thought. The atmosphere is trapping more heat. More species are threatened. Ever more disastrous outcomes will lead to even greater warming.

But two Los Alamos scientists, working with an international team from other government laboratories and universities, recently discovered that at least one aspect of the warming climate is actually *less* of a concern than previously believed. Stephen Price and Matthew Hoffman of Los Alamos, working with Mauro Perego of Sandia National Laboratories, carried out supercomputer simulations of the Greenland ice sheet with two Department of Energy-supported models. The simulations were based upon an understanding derived from recent field measurements of the Greenland ice sheet carried out by other members of the international collaboration. Taken together with contributions from two European models, the simulations showed that future increases in meltwater running beneath the ice sheet will have a smaller-than-expected influence on ushering glacial ice into the ocean.

For the past decade, glaciologists have debated whether or not such meltwater could accelerate the demise of the Greenland ice sheet and therefore the pace of sea level rise overall. Meltwater flowing along the surface ice can dive into moulins—vertical shafts that convey water to the base of the ice sheet, where it spreads across the underlying bed. As the conventional thinking goes, this water ought to lubricate the interface between the ice and the ground, thereby causing glaciers to flow more quickly into the ocean. But just how significant this lubricating effect ought to be has been an issue of contention.

Now, for the first time, detailed, credible predictions of the contribution of meltwater lubrication to sea level rise are available. The research revealed that by the end of the century the effect will add at most 4 percent to the overall sea level rise from Greenland.

"By the year 2100, Greenland's contribution to global sea level rise is projected to be about 6 centimeters, with the majority of that attributable to increased melting alone," Price says, referring to results based on a probable greenhouse gas emissions scenario during that time period. "But the additional sea level rise that's caused by meltwater lubrication will be only a few millimeters." Price says that by 2200, Greenland's estimated contribution to sea level rise from melting should be about 17 centimeters, with less than one additional centimeter due to the meltwater lubrication effect. And sometimes, increased meltwater can actually inhibit the flow of a glacier.

As the Greenland summer presses on, larger flows of meltwater open up progressively wider tunnels at the bottom of the ice. When summer ends and less meltwater flows through the oversized tunnels, the water is able to drain more efficiently without lubricating the ice-rock interface, thereby reducing the motion of the ice.

When winter arrives, the mass of the glacier crushes down into the empty meltwater channels, such that when the flows resume in the spring, the underlying drainage network is insufficient to accommodate all the water. The back-pressure of the non-draining water tends to lift the glacier, reinstating the lubrication effect and helping the glacier move seaward. The lubrication effect persists until the underlying water channels once again outgrow what the flows require, allowing the water to drain without lifting the glacier. The new computer simulation accounts for both annual effects on the ice flow—slowing in the fall and accelerating in the spring—and yields a compromise outcome in which the meltwater from moulins only minimally hastens the movement of ice into the sea.

Should everyone breathe a collective sigh of relief? Maybe just a little. As Hoffman says, "I find it reassuring to put a box around this particular process and show it to be not as serious as was once thought. This allows us to turn our attention to better understanding other potential contributions to sea level rise from glaciers and ice sheets." LDRD

Meltwater on the Greenland ice sheet forms surface flows that can eventually dive into moulins. These moulins deliver water to the base of the ice sheet, where it marginally affects the rate at which glaciers move to the sea.
CREDIT: MATTHEW HOFFMAN/LANL

How to Spot a Nuke

The United States needs a fast, reliable screening system to scan for nuclear materials like uranium and plutonium at ports, borders, and other sensitive areas. Los Alamos National Laboratory recently demonstrated a new technology that does exactly that.

Andrea Favalli and Martyn Swinhoe of the Lab's Nuclear Safeguards Science and Technology group led an experiment in which

they were able to detect and measure the quantity of nuclear material inside a closed container, using a laser pulse lasting less than a trillionth of a second. During that instant, the laser delivers 50 times more power than the rest of the world's electrical power production combined.

"This kind of laser-driven nuclear-material detection was just an idea," says Favalli. "No one knew if it could actually be done until we worked out the details, fabricated the parts, and performed the test."

Favalli, Swinhoe, and their team focused the Lab's powerful TRIDENT laser onto a thin plastic target, concentrating the incredibly high-energy burst of light into a spot less than a thousandth the diameter of a human hair. The plastic had been previously deuterated, meaning that its hydrogen atoms were replaced with deuterium, a heavier isotope of hydrogen with a loosely bound proton-neutron pair comprising its nucleus. The laser blasts the deuterium nuclei off the plastic in a high-speed beam that strikes a second target made of metal. When the deuterium nuclei strike the metal target, they split apart and shake loose a tremendous, billionth-of-a-second shower of neutrons traveling at up to half the speed of light. These high-energy neutrons, originating from both the deuterium and the metal nuclei, penetrate the closed container being scanned.

Normally, the neutron burst would be detected after passing through the container and that's the end of the story. However, when nuclear materials are present, the neutron burst will cause some nuclear fission reactions within the material. (Such nuclear material is always in a noncritical configuration unless deliberately detonated in a bomb, so these additional fissions pose no danger; a complete nuclear weapon can be safely scanned in this way.) The fissions produce a wave of additional neutrons, called delayed neutrons, which can be detected for several seconds after the laser-driven burst. It is these delayed neutrons that reveal the presence of illicit nuclear materials.

Laser experts within the research group are confident that the entire laser-driven neutron detection system can be shrunk down to fit within the back of a shipping truck, making it portable enough to distribute to border points and other locations where needed. The technique may also find applications in scientific research as a convenient neutron source for studying the effects of radiation on materials and electronic systems, among other uses. LDRD

Explosives Going Dark

In addition to developing field-deployable technology for detecting nuclear materials [see previous Spotlight article], Los Alamos also contributes to field-deployable technology for detecting conventional explosives. Research carried out by a team from the U.S. Air Force Academy recently showed that an enhanced biomarker, developed at Los Alamos, can rapidly screen for certain dangerous explosives and toxins—to the benefit of military and civilian security officers, first responders, and humanitarian remediation workers.

The biomarker is a type of green fluorescent protein, or GFP. Normally used in biological research, GFP emits a characteristic green glow when exposed to blue light, making it easy to detect cellular components tagged with the marker. GFP can also be triggered to glow green by exposure to 280-nanometer wavelength ultraviolet (UV) light—a feature that has been largely ignored because UV light causes cellular damage. However, in a non-biological context, this feature can be exploited to indicate the presence of nitroorganic high explosives, including TNT and RDX. The explosives inhibit the UV-excitation mechanism, so when they are added to UV-illuminated GFP, the green glow shuts off—an event easily recognized in a laboratory or field setting.

Los Alamos biologists Andrew Bradbury, Geoff Waldo, and Csaba Kiss created an enhanced version of GFP, called eCGP123 ("enhanced consensus green protein"), capable of withstanding the rigors of field-deployment. Unlike other materials-sensing molecules, which can be rendered ineffective by exposure to common chemicals or elevated temperatures, eCGP123 is highly stable—and it responds strongly to each of the six explosives tested. It can be produced inexpensively in large quantities and may even be reusable as well: an hour after being exposed to explosives in vapor form, eCGP123 resumed its green glow.

Early indications suggest that the class of organic materials capable of inhibiting the UV excitation of GFP likely includes not only explosives, but a number of poisons and chemical-weapon nerve agents, too. LDRD

—Craig Tyler

Thermostability galore: fluorescent protein eCGP123 continues to glow even when briefly boiled.

ISSN: 1942-6631

Address mail to:
1663
Mail Stop M711
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
1663magazine@lanl.gov
www.lanl.gov/newsroom/publications/1663/

Presorted Standard
U.S. Postage Paid
Albuquerque, NM
Permit No. 532

Autumn colors along the Rio Chama in Northern New Mexico.

Los Alamos NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396.
A U.S. Department of Energy Laboratory LALP-13-003

Printed on recycled paper